Sondage ESPRI-IA

Machine Learning dans les laboratoires de l'IPSL

Sébastien Gardoll

2 octobre 2020

Table des matières

1 Sondage

1.1 Préambule

Ce rapport présente une analyse du sondage mené auprès des membres des laboratoires de l'IPSL au sujet de leurs attentes et pratiques en matière de Machine Learning. Le sondage a été présenté lors du webinaire « Cluster GPU et applications scientifiques », le 26 mai 2020 et 62 personnes y ont répondu jusqu'à présent. Il a été conçu et analysé par Yann Delcambre (LATMOS), Sébastien Gardoll (IPSL), Maya George (LATMOS), Cécile Mallet (LATMOS).

1.2 Questions

Le sondage a été implémenté à l'aide de Google Forms. Il est composé de questions à choix multiples et de questions à réponse librement rédigée par les sondés. Pour les questions à choix multiples, on distingue les questions n'acceptant qu'une seule réponse parmi celles proposées, des questions acceptant plusieurs réponses. Pour chacune des questions, ce document précise leur type entre crochets : unique, multiple ou libre. À noter que les questions de type multiple proposent toujours aux sondés de rajouter une nouvelle réponse (option « autres »). Les questions de type unique proposent un ensemble de réponses couvrant tous les cas de figure. Hormis la question relative au rattachement au laboratoire, les réponses aux questions ne sont pas obligatoires. Le lecteur de ce document doit toujours considérer le nombre de sondés ayant répondu à chacune des questions. Voici la liste des 16 questions du sondage :

1. À quel laboratoire êtes vous rattaché·e ? [unique]
2. Quels sont vos thèmes scientifiques et/ou vos domaines d'ingénierie, concernés par le Machine Learning ? [libre]
3. Quel est votre degré de compétences en Machine Learning ? [unique]
4. Quel est le degré de maturité de vos projets Machine Learning ? [unique]
5. Quelles sont les origines de vos données concernées par le Machine Learning ? [libre]
6. Avez vous des partenariats avec des laboratoires extérieurs à l'IPSL ou des entreprises, concernant les aspects Machine Learning de vos projets ? [libre]
7. Quelles sont vos attentes ou remarques au sujet d'ESPRI-IA ? [libre]

8. Si vous souhaitez être contacté·e, laissez nous votre email à cet endroit (optionnel) [libre]
9. Quelles méthodes ML avez vous déjà mises en œuvre (plusieurs choix possibles) ? [multiple]
10. Quels objectifs avaient vos modèles ML (plusieurs choix possibles) ? [multiple]
11. Quelles plateformes de calculs avez vous utilisées pour vos projets ML (plusieurs choix possibles) ? [multiple]
12. Avez vous utilisé une plateforme GPU pour vos projets ML (plusieurs choix possibles) ? [multiple]
13. Quels sont les langages informatiques, bibliothèques et plateformes d'exécution (CPU et/ou GPU) utilisés pour vos projets ML ? Si le langage est compilé (C, Fortran, etc.), veuillez indiquer le nom du compilateur. [libre]
14. Quels outils de programmation avez vous utilisés pour réaliser vos projets ML (plusieurs choix possibles) ? [multiple]
15. Avez vous eu des problèmes pour traiter vos données pour vos projets Machine Learning (plusieurs choix possibles) ? [multiple]
16. Avez vous des commentaires ? [libre]

Pour des raisons de confidentialité résumées à cette adresse, il n'a pas été demandé aux sondés des informations à caractère personnel (identité, statuts, etc.).

1.3 Transformations

Google Forms propose d'exporter les réponses aux questions sous forme de fichier CSV (*ESPRI-IA ML sondage.csv*). Les colonnes représentent les questions et les lignes les réponses aux questions d'un sondé. Afin de réduire l'encombrement des graphiques, les questions et les réponses proposées sont condensées en mots clefs dont les correspondances sont données en annexe A.1. Par exemple, « À quel laboratoire êtes vous rattaché·e ? » devient « lab ». Les questions suivantes, à réponse libre, ont fait l'objet d'une interprétation afin de réduire leurs réponses à un ensemble de mots clefs assimilables à des catégories :

- Quels sont vos thèmes scientifiques et/ou vos domaines d'ingénierie, concernés par le Machine Learning ?
- Quelles sont les origines de vos données concernées par le Machine Learning ?
- Avez vous des partenariats avec des laboratoires extérieurs à l'IPSL ou des entreprises, concernant les aspects Machine Learning de vos projets ?
- Quelles sont vos attentes ou remarques au sujet d'ESPRI-IA ?

- Quels sont les langages informatiques, bibliothèques et plateformes d'exécution (CPU et/ou GPU) utilisés pour vos projets ML ?

Les interprétations figurent dans le fichier *analyse_champs_libre.ods* et sont ajoutées aux données par un script (*concatenation.py*) pour former le nouveau fichier de données : *extended_sondage.csv*.

1.4 Graphiques

Les graphiques de ce document ont été générés par le script *sondage.py* implémenté en Python3 (voir annexe A.2 pour les dépendances). La nature des réponses aux questions de type unique et multiple étant catégorielle, la production de leurs graphiques unidimensionnelles (répartition des réponses relatives à une seule question ; diagrammes à barres) n'a pas nécessité de traitements de données. Pour la production de leurs graphiques bidimensionnelles (corrélation entre les réponses de deux questions), on distingue deux cas : le cas entre deux questions de type multiple et les autres combinaisons. Pour le premier cas, j'ai décidé de rechercher des regroupements de sondés à l'aide de la méthode kmeans (dendrogrammes pour le contrôle du nombre de groupes) et d'illustrer ces groupes par projections ACP des données (nuages de points colorés selon les groupes). Les données sont encodées par la méthode *one hot* au préalable. Les caractéristiques des centroïdes des groupes trouvés sont enregistrées dans des fichiers préfixés *kmeans*. Pour le deuxième cas, la production de graphiques bidimensionnelles (carte de fréquentation ou *heat map*) n'a besoin que d'un encodage *one hot* puis de la construction d'une table de contingence.

1.5 Mises à jour

Tous les fichiers (ce rapport, les scripts, graphiques, données, etc.) sont accessibles à l'adresse suivante : https://gitlab.in2p3.fr/ipsl/espri/espri-ia/sondage_2020

2 Analyses

2.1 Participation

La figure 1 donne la répartition des sondés selon leur laboratoire de rattachement (unique). Cette question est la seule obligatoire et nous permet donc de connaître l'origine de tous les sondés (62). On constate que la majorité des sondés proviennent du LATMOS (40,3 %) puis du LMD et du LSCE (les deux en-

viron 16 %). Les laboratoires ne sont pas représentés de façon équilibrée. Donc les analyses mettant en corrélation les laboratoires doivent fortement prendre en compte le nombre de sondés par laboratoire. Pour la suite de ce document, le lecteur devra garder en mémoire que les analyses ou représentations bidimensionnelles se basant sur le rattachement des sondés sont fortement biaisées par la contribution du LATMOS, LMD et LSCE. On évitera de comparer les distributions entre deux laboratoires et on considérera uniquement la distribution au sein d'un laboratoire.

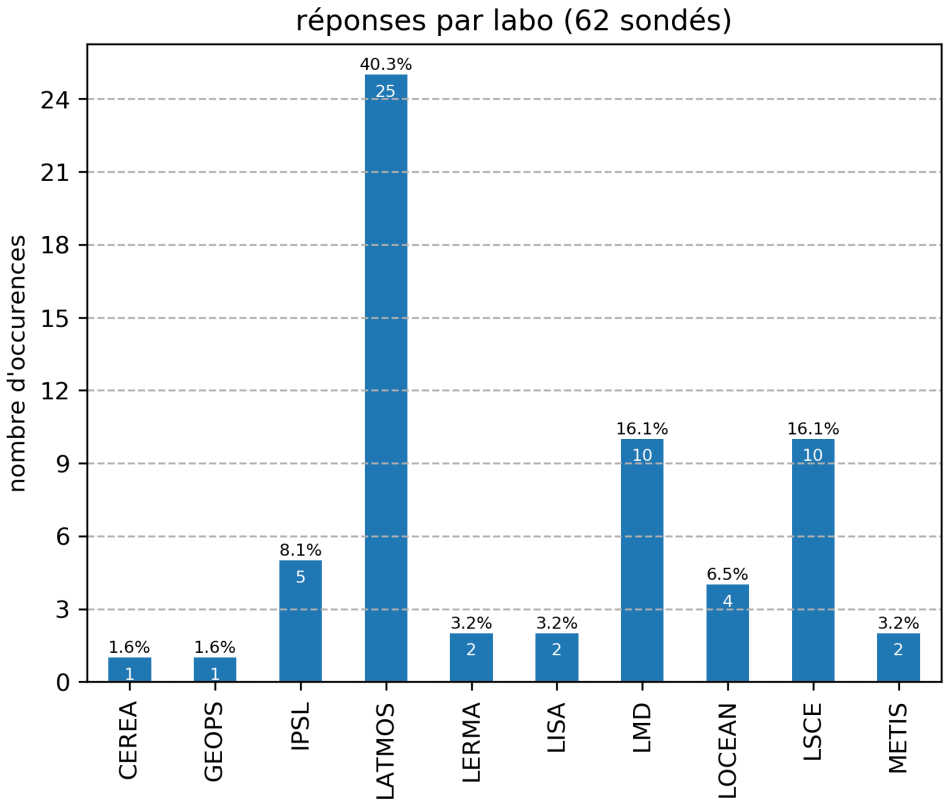


Figure 1 : nombre de sondés par laboratoire

2.2 Thèmes scientifiques et techniques

La figure 2 présente une interprétation de la réponse libre de 56 sondés concernant leurs thèmes scientifiques et techniques en rapport avec le Machine Learning (libre). À parts égales les thèmes atmosphère, simulations climatiques et télédétection, représentent les thèmes majoritaires des sondés. Ces thèmes reflètent la distribution des sondés selon leur laboratoire (test du χ^2 est positif au risque α de 0,1 % avec un V de Cramér à 0,35 ; voir annexe 21), heureusement ! Le deuxième groupe en importance est composé des thèmes événements extrêmes, modélisation statistique et inverse. À noter que l'on retrouve quasiment tous les domaines scientifiques couverts par l'IPSL. Pour information, on

distingue modélisation statistique du Machine/Deep Learning. Le premier est une application du Machine/Deep Learning à la modélisation physique et le deuxième est la recherche dédiée au Machine/Deep Learning.

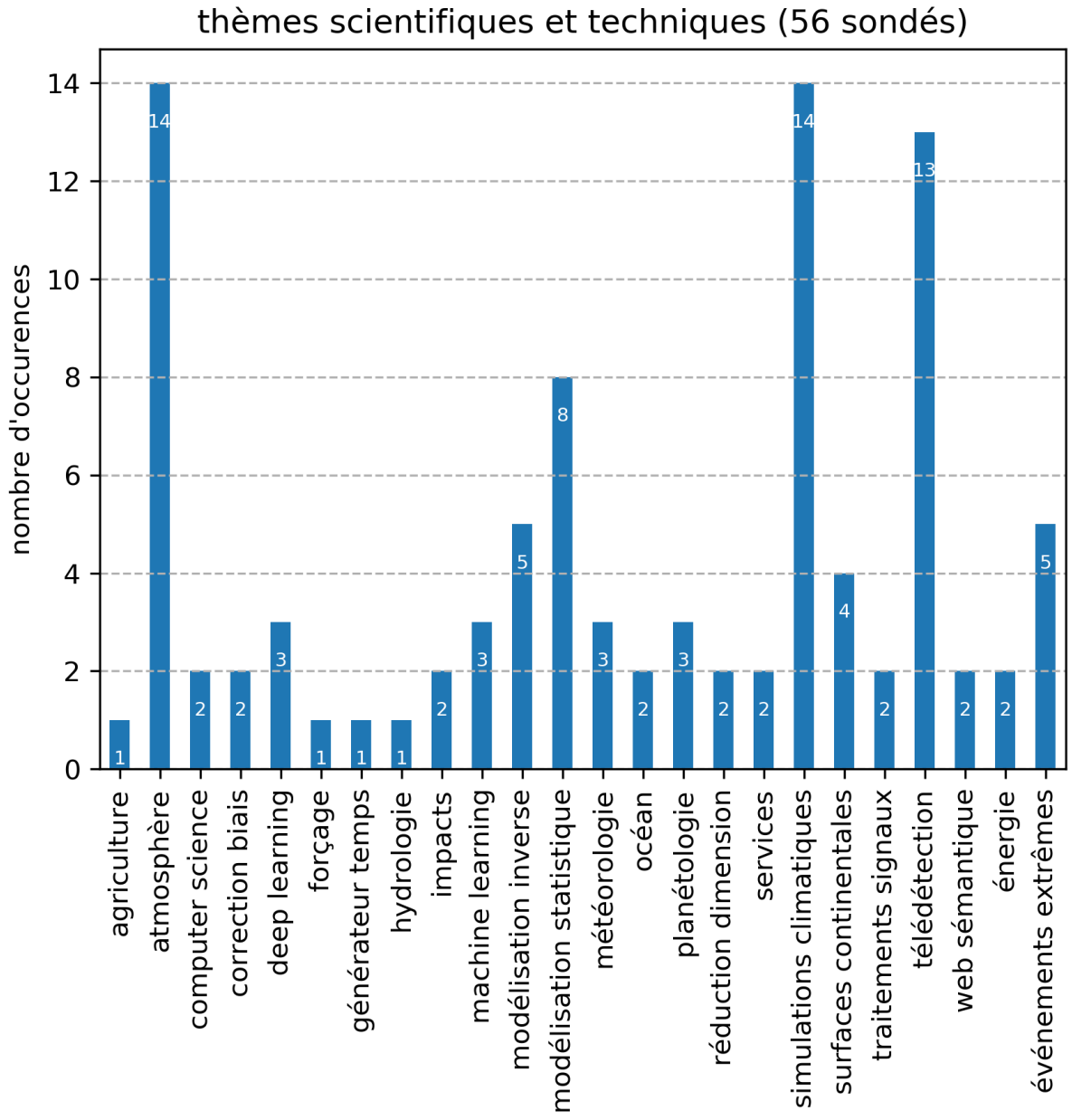


Figure 2 : répartition des thèmes scientifiques et techniques

2.3 Compétences

La figure 3 illustre la répartition du degré de compétence en Machine Learning (unique) de tous les sondés (62). On constate que la moitié des sondés n'a pas de compétences en ML, mais a un certain intérêt pour le ML. L'autre moitié a des compétences, mais 17,7 % des sondés ne sont pas autonomes. En annexe, la figure 22 donne la répartition par laboratoire (voir avertissement section 2.1).

La figure 4 nous renseigne que les thèmes des simulations climatiques, de l'atmosphère et de la télédétection regroupent la majorité des sondés intéressés par le ML.

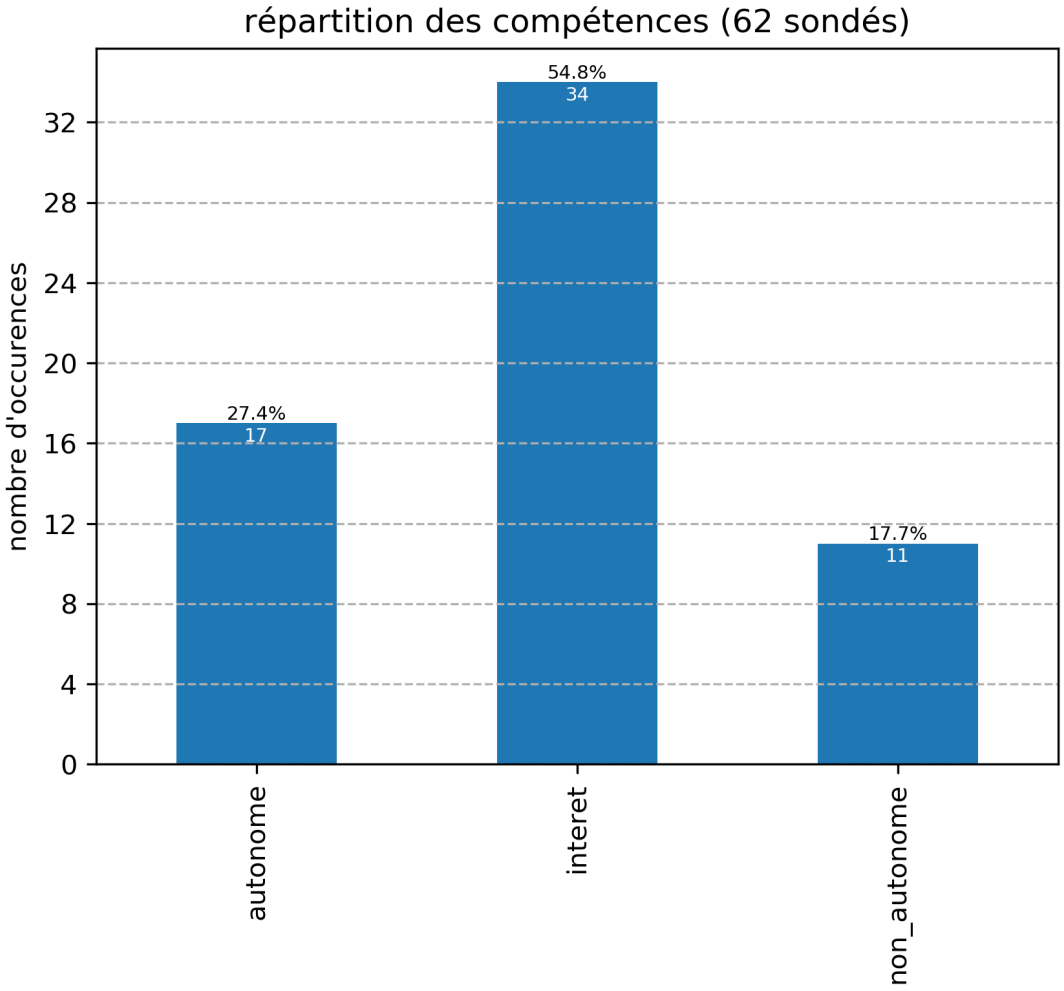


Figure 3 : répartition des compétences des sondés

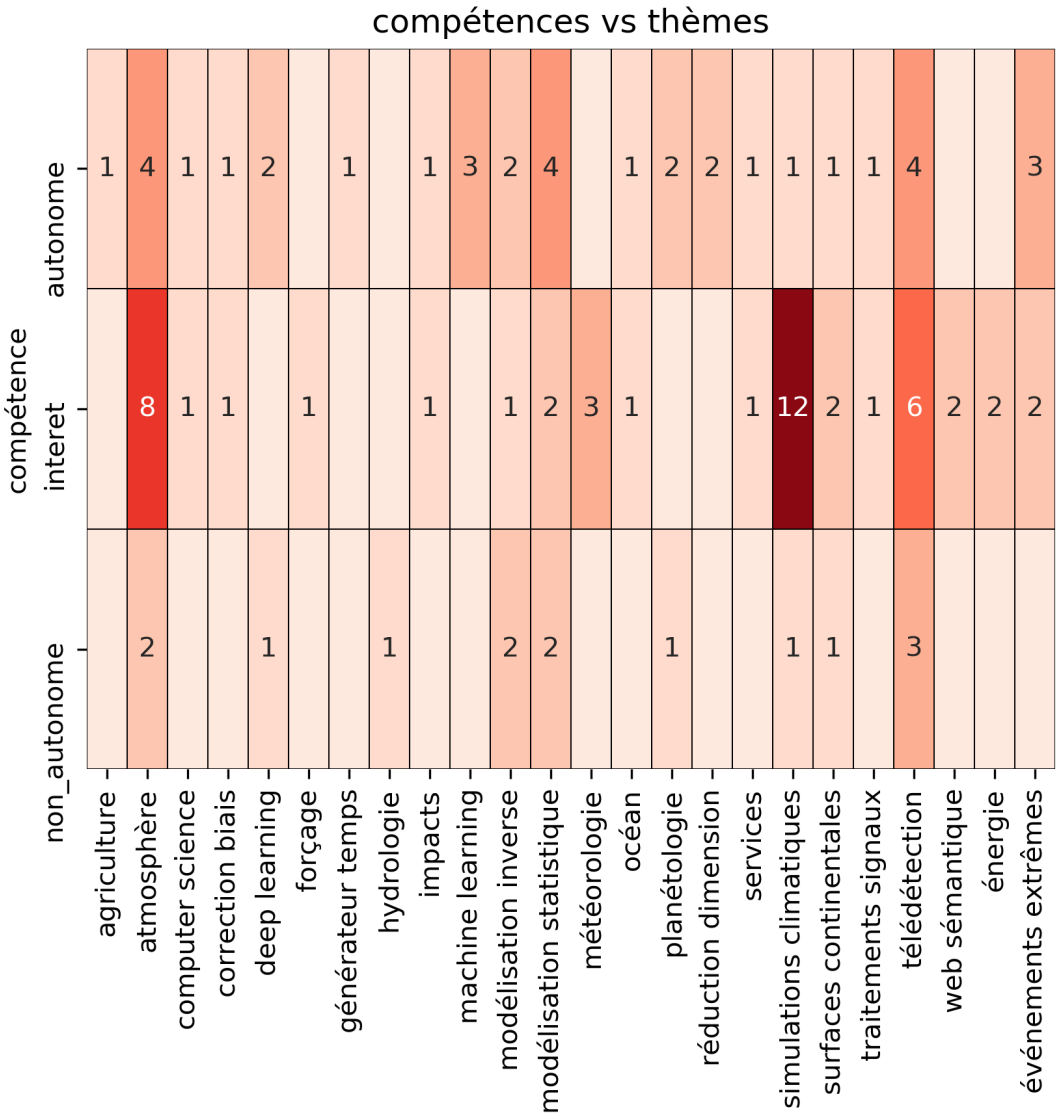


Figure 4 : compétences des sondés par thèmes scientifiques et techniques

2.4 Maturité des projets

Le degré de maturité des projets ML (unique) de tous sondés est représenté à la figure 5. La moitié des sondés déclare avoir fait ou avoir piloté des projets ML, avec 17,7 % des projets mettant en œuvre des méthodes ML spécifiquement adaptées. 30 % des sondés ont une idée de projet ML et 15 % n'en ont pas mais pensent que leurs données sont valorisables. Les figures 23 et 24 en annexe donnent la répartition du degré de maturité des projets, respectivement par laboratoire et par thèmes (voir avertissement section 2.1).

Bien sûr, degré de maturité des projets et degré de compétence des sondés sont corrélés (test du χ^2 est positif au risque α de 0,1 %, avec un V de Cramér à 0,41). La figure 6 représente sa carte de fréquentation. Elle révèle qu'au moins 18 personnes des laboratoires de l'IPSL ont besoin d'aide afin de démarrer un

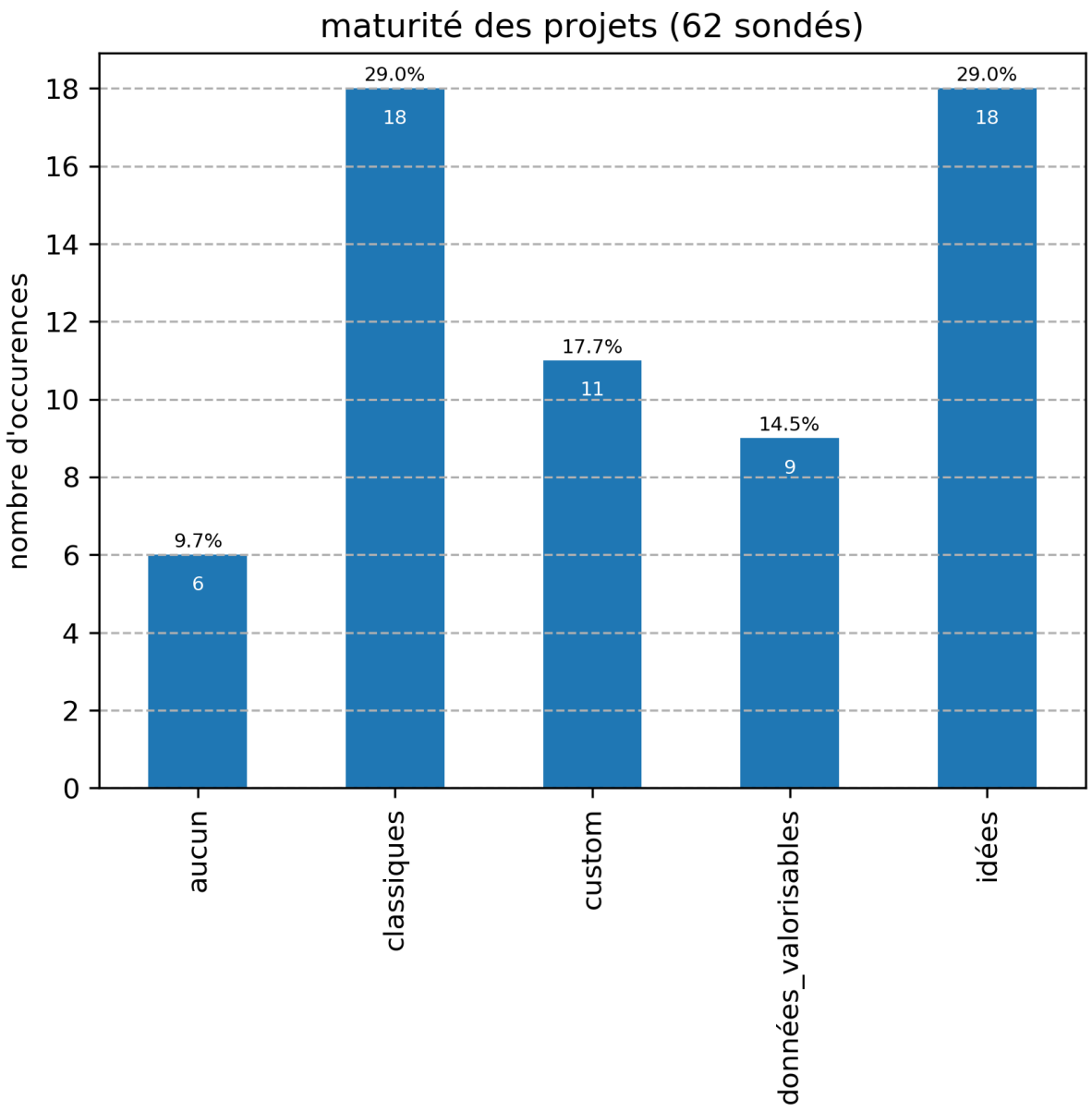


Figure 5 : degré de maturité des projets ML des sondés

projet et 12 personnes ont besoin de conseils afin de réaliser leurs projets. ESPRI-IA trouve donc une justification pour au moins 30 sondés.

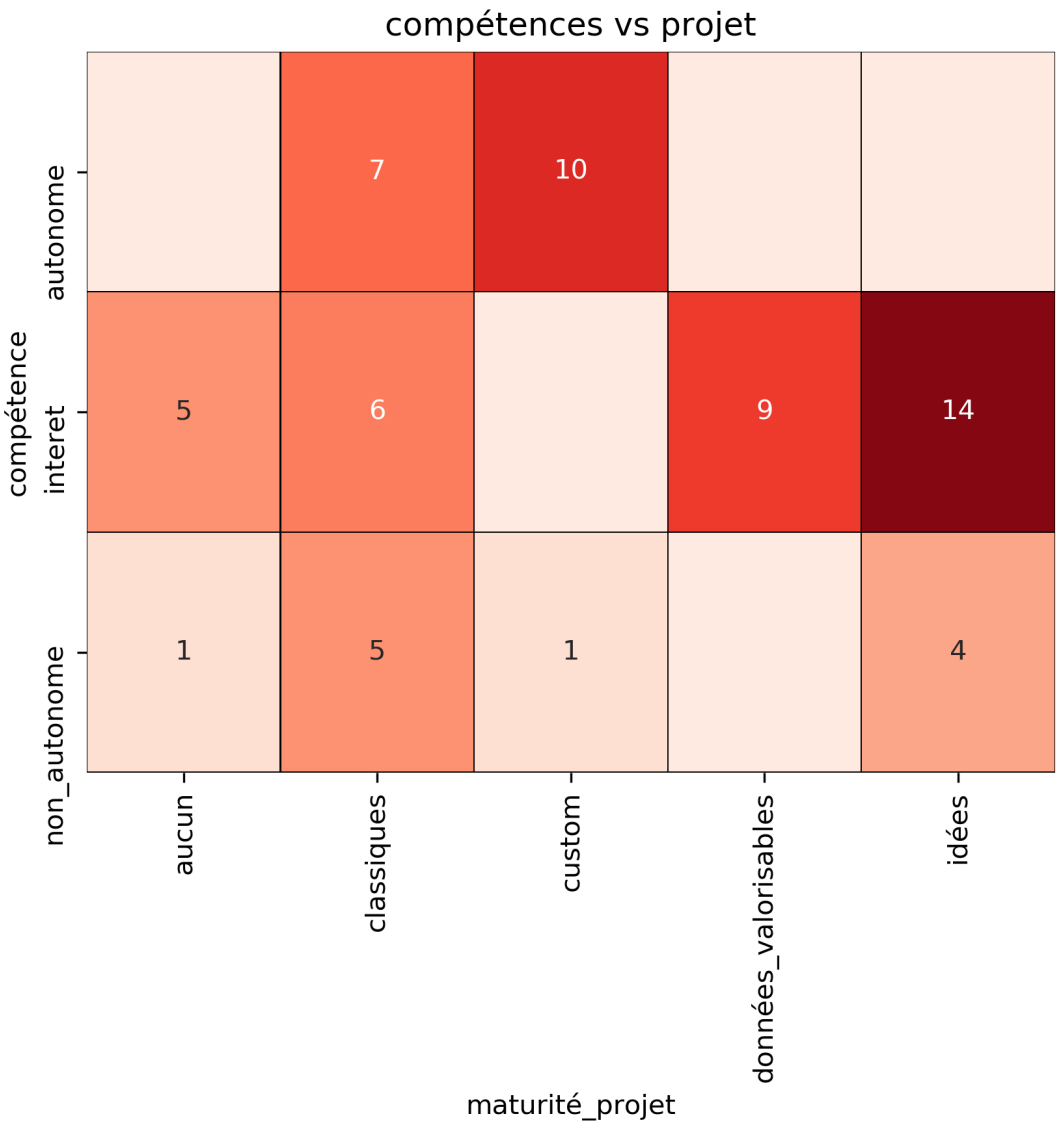


Figure 6 : degré de compétence des sondés versus degré de maturité de leurs projets

2.5 Sources des données

Selon la figure 7, les données concernées par le ML (libre) sont pour 52 sondés stockées majoritairement sur le mésocentre de l'IPSL (Ciclad et Climserv) et sur leur machine personnelle. L'origine des données est assez bien étalée mais il semble qu'une partie provienne préférentiellement des expériences CMIP5&6, de ERA5, de SIRTA, de la NASA et de IASI.

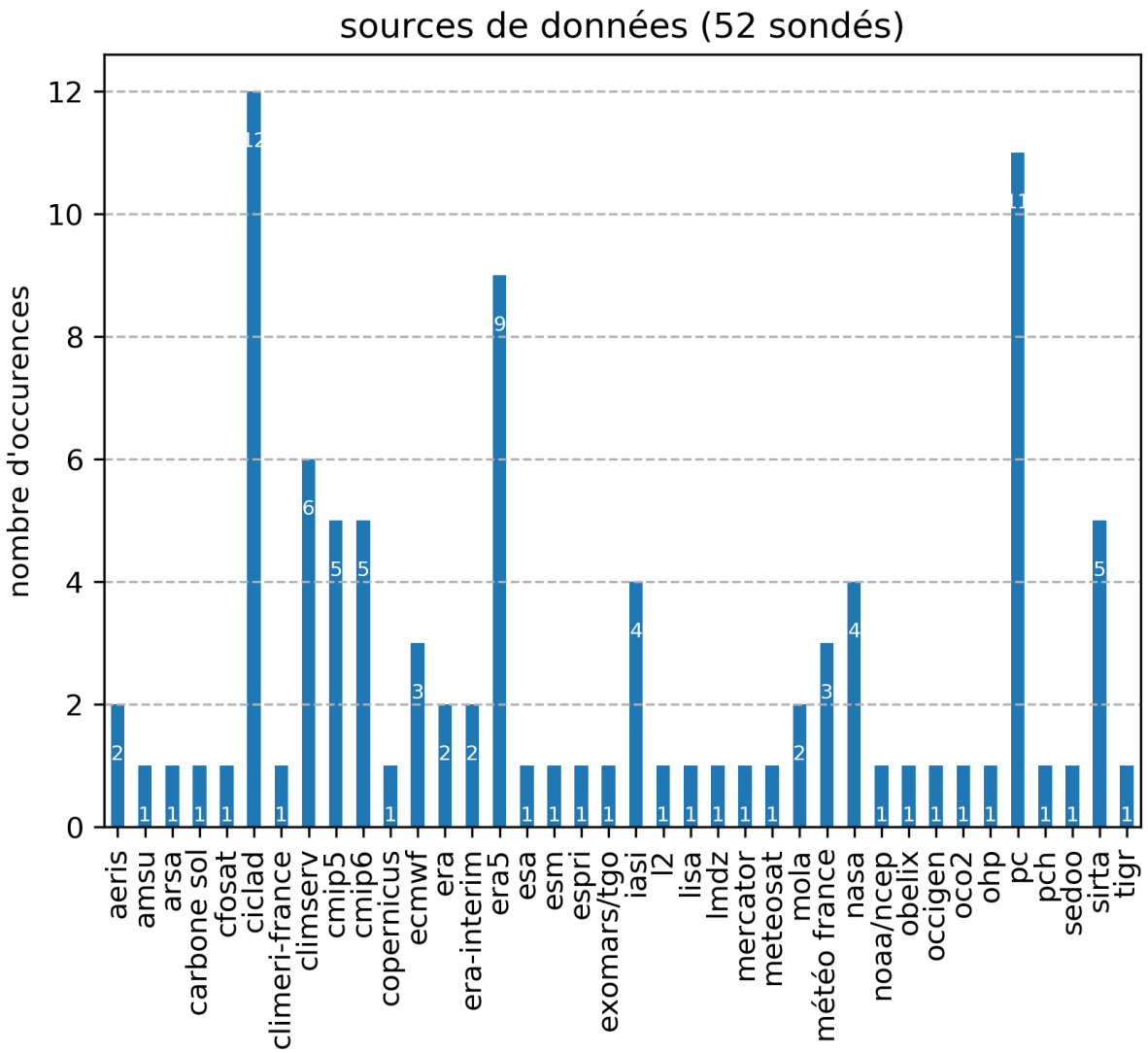


Figure 7 : origines et stockages des données concernées par le ML

2.6 Partenariats

13 sondés sont impliqués dans un ou plusieurs partenariats ayant un rapport avec le ML (libre). La figure 8 donne une répartition des partenariats de ces sondés par laboratoire (voir avertissement section 2.1).

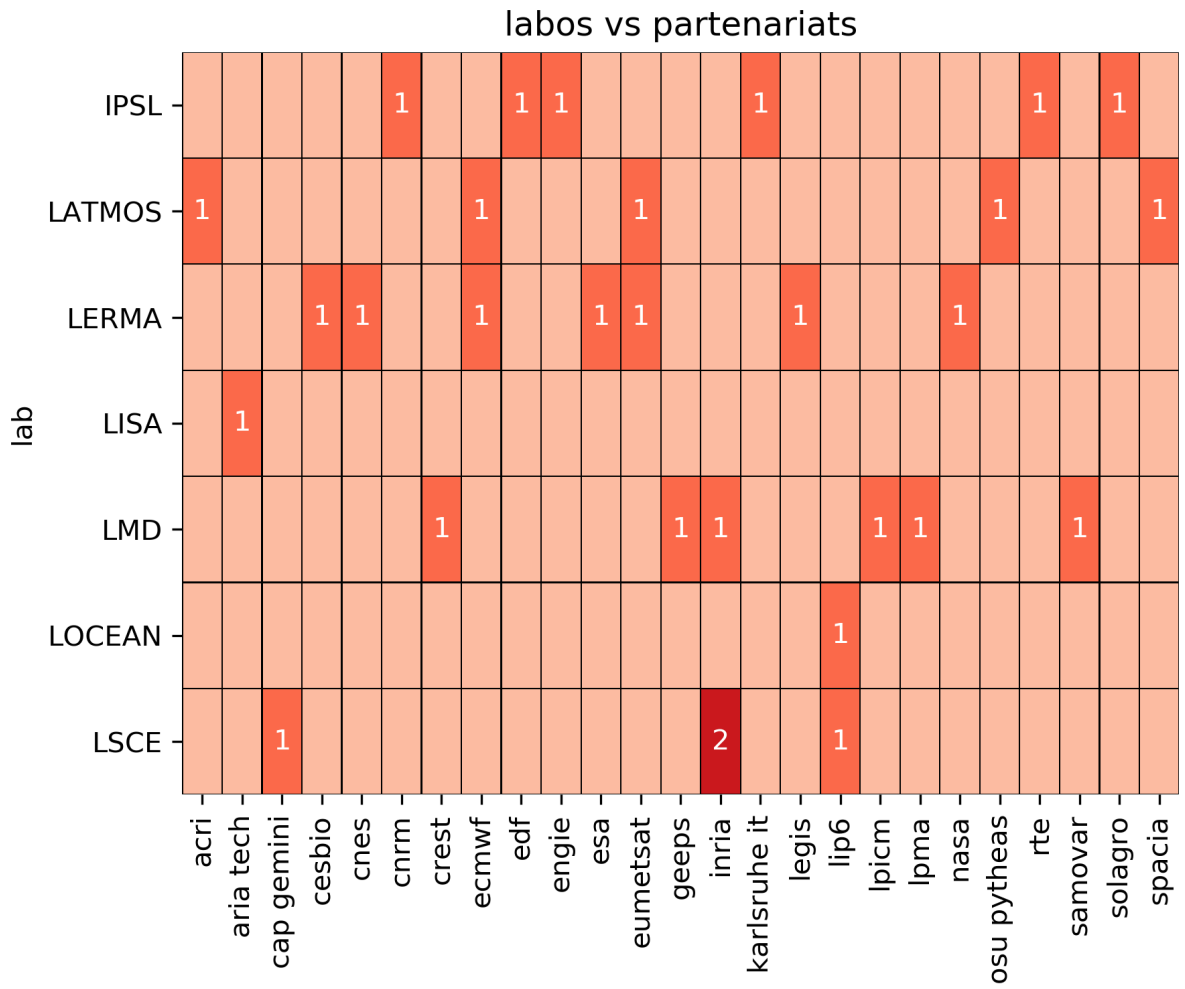


Figure 8 : Les partenariats de 13 sondés par laboratoire

2.7 Méthodes de Machine Learning

La distribution des méthodes ML utilisées (multiple) par 44 sondés est représentée sur la figure 9. On constate que toutes les méthodes sont utilisées, avec une majorité de régressions polynomiales (non neuronales) à 18,5 % suivies par les méthodes de clustering Kmeans à 13,7 %. Le perceptron multicouche (MLP), une alternative neuronale aux régressions polynomiales (mais aussi aux méthodes de classification), est à 11,3 %. Les cartes auto-organisatrices (SOM), alternatives neuronales à Kmeans et à l'ACP (méthodes factorielles), sont à 6,5 %. Les méthodes de Deep Learning (DL) ne représentent que 8,9 %. Si l'on cumule les méthodes neuronales (SOM, DL et MLP), elles représentent 27,3 %. Le lecteur de ce document ne devrait pas tirer de conclusions sur l'adéquation des méthodes neuronales à partir du nombre de sondés les ayant utilisées, non plus au sujet de leur diffusion dans les laboratoires de l'IPSL. La distribution des méthodes par laboratoires est donnée en annexe 25 (voir avertissement section 2.1).

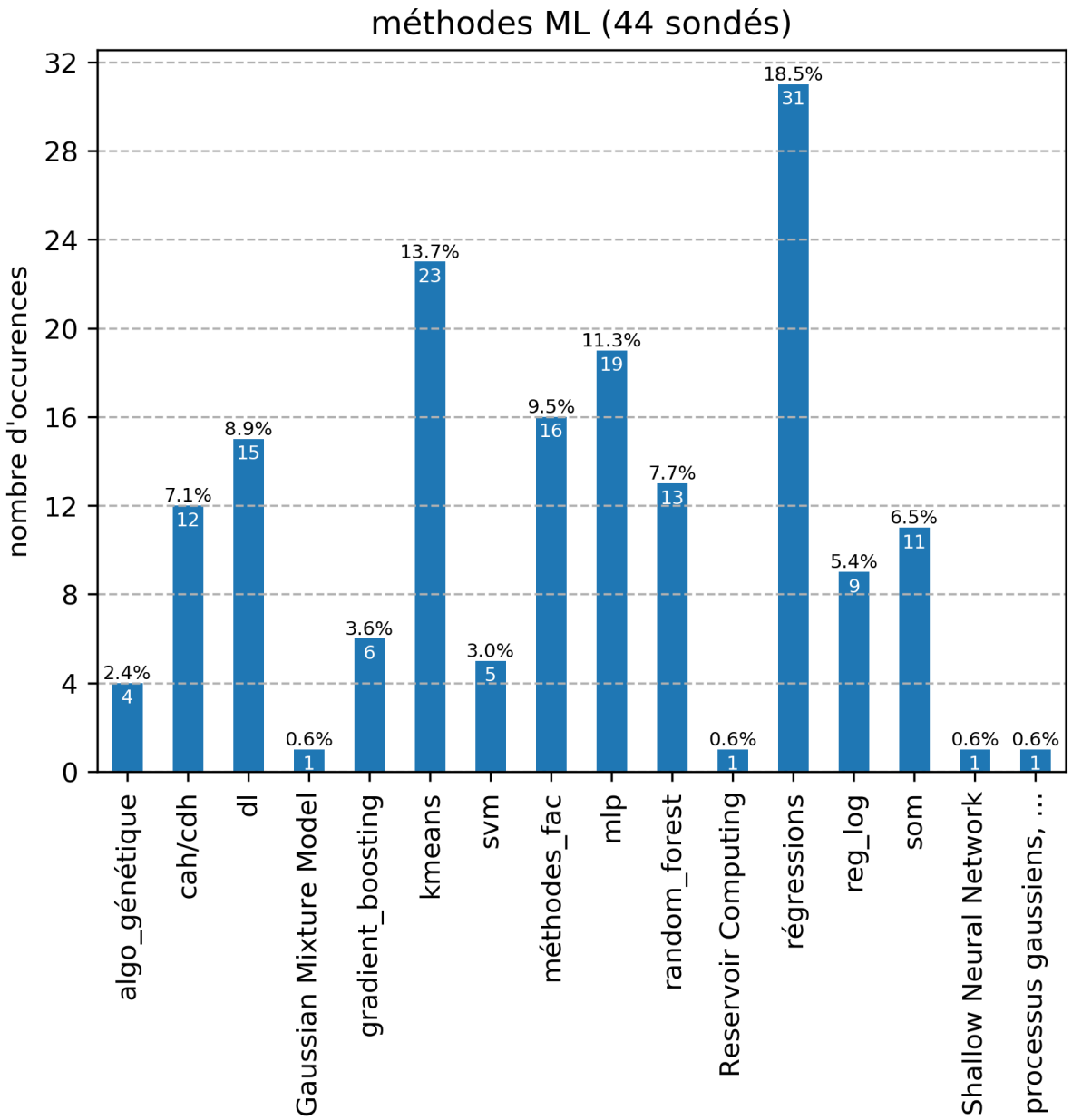


Figure 9 : répartition de l'utilisation des méthodes ML

La carte de fréquentation 10 met en relief les compétences des sondés (44) vis à vis des méthodes ML utilisées. On s'attend à ce que compétences et méthodes soient corrélées, mais le test du χ^2 montre que les deux variables ne le sont pas dans le cadre de cette étude (risque α très supérieur à 5 %). Si l'on regarde la ligne des sondés ayant un intérêt pour le ML sans en avoir de compétences, il est curieux de voir autant de pratiques du Deep Learning et du MLP, des méthodes techniquement pointues (sans parler des autres méthodes). Peut-on considérer que ces sondés ont piloté des projets ML utilisant ces méthodes ?

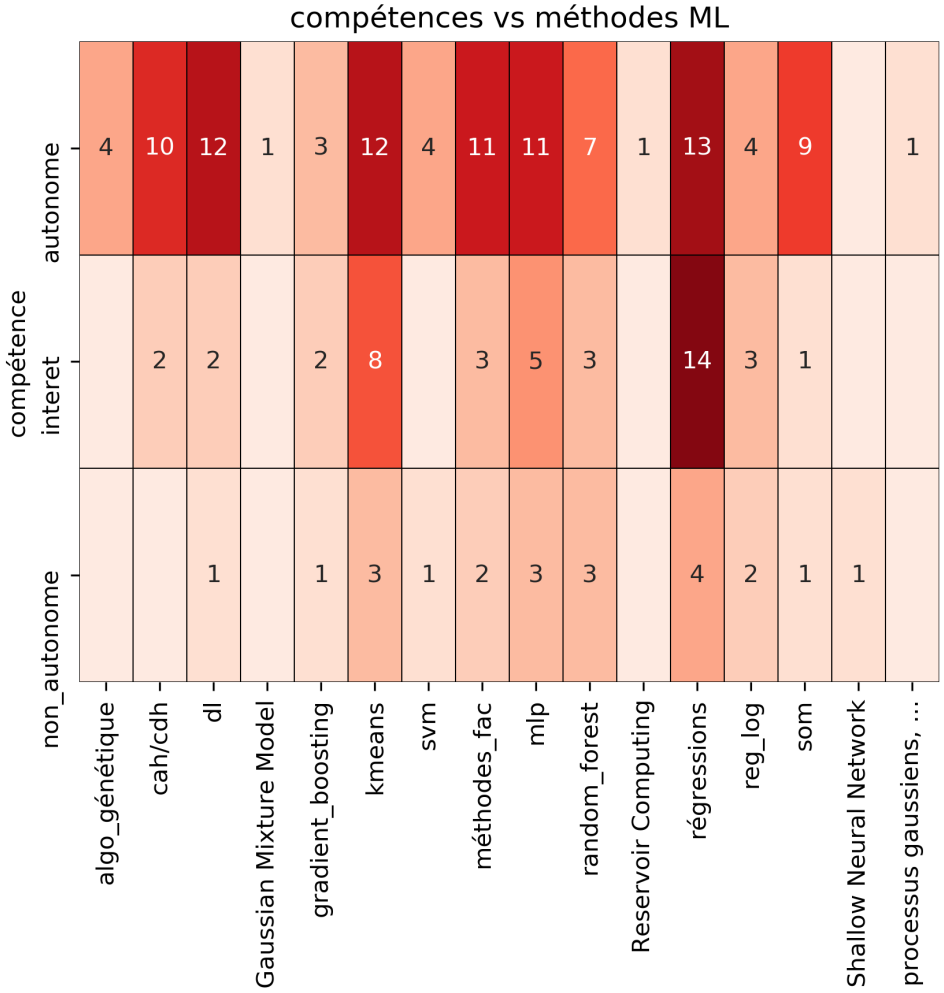


Figure 10 : relation entre compétences des sondés et utilisation de méthodes ML

2.8 Usages du Machine Learning

La figure 11 présente la répartition des usages du ML pour 50 sondés. On s'attend à retrouver les reflets des méthodes ML majoritaires, ce qui est le cas pour la régression (17,1 %) et le clustering (11,2 %). La classification, deuxième usage du ML atteint 16,4 %. Ce nombre s'explique par l'usage des différentes méthodes de DL/MLP, arbres de décisions (Random Forest et Gradient Boos-

ting) et machines à vecteurs de support (SVM) qui sont capables de régressions comme de classifications. La répartition est assez homogène mise à part la génération de données et la segmentation, deux usages de niches impliquant des architectures DL complexes.

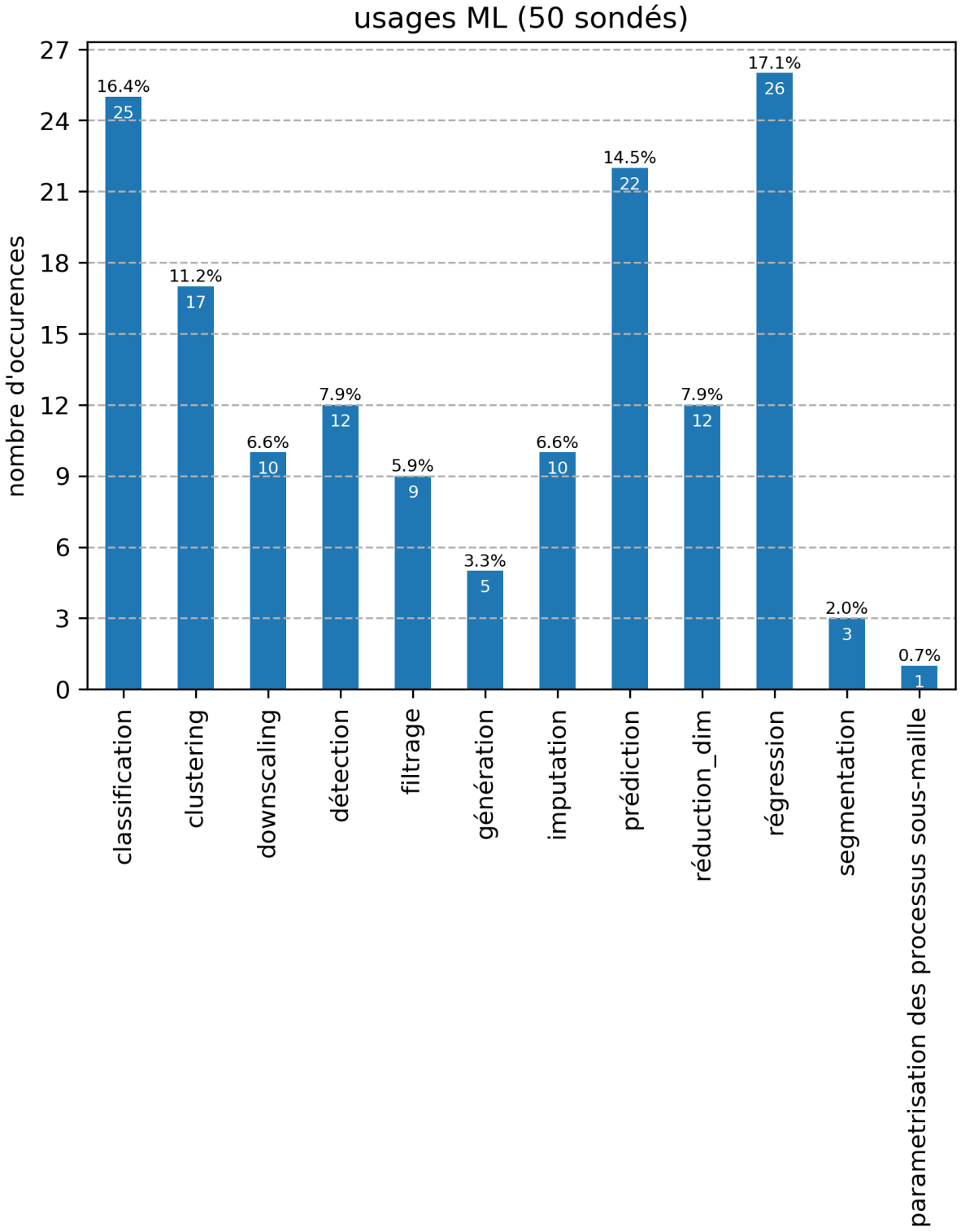


Figure 11 : répartition des usages du ML

Les figures 26 et 27 en annexe, donnent respectivement les usages par laboratoires (voir avertissement section 2.1) et par compétences des sondés. Usages et laboratoires ne sont pas corrélés (test du χ^2). Idem entre usages et compétences, cependant, cette relation est biaisée comme l'est la relation entre méthodes ML et compétences (voir section 2.7).

L'analyse de la relation entre méthodes et usages ML (toutes les deux de type multiple) nous apprend qu'il existe trois groupes de sondés (voir *kmeans_centres_methodes_usages.csv*) :

- 17 sondés utilisant des méthodes ML non neuronales principalement pour la régression et dans une moindre mesure la classification et le clustering — le groupe des régresseurs —
- 16 sondés utilisant presque uniquement les méthodes neuronales (sauf SOM) principalement pour la classification, la prédiction et le downscaling — le groupe des réseaux neuronaux uniquement —
- 11 sondés utilisant toutes les méthodes pour tous les usages — le groupe du ML complet —

Les trois groupes sont illustrés à la figure 12 : une projection sur les deux premières composantes principales des deux variables (ACP). Les dendrogrammes de contrôle du nombre de groupes sont donnés en annexe 28 et 29.

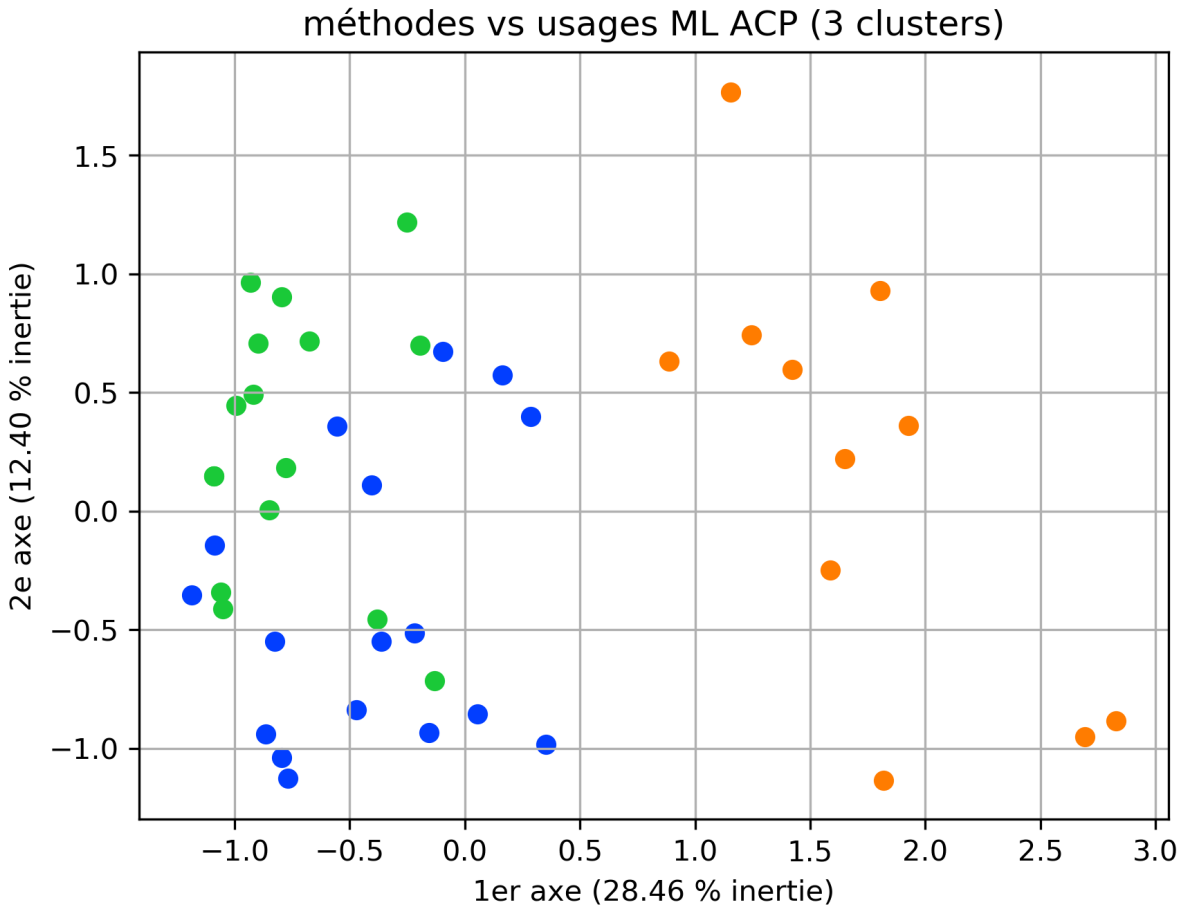


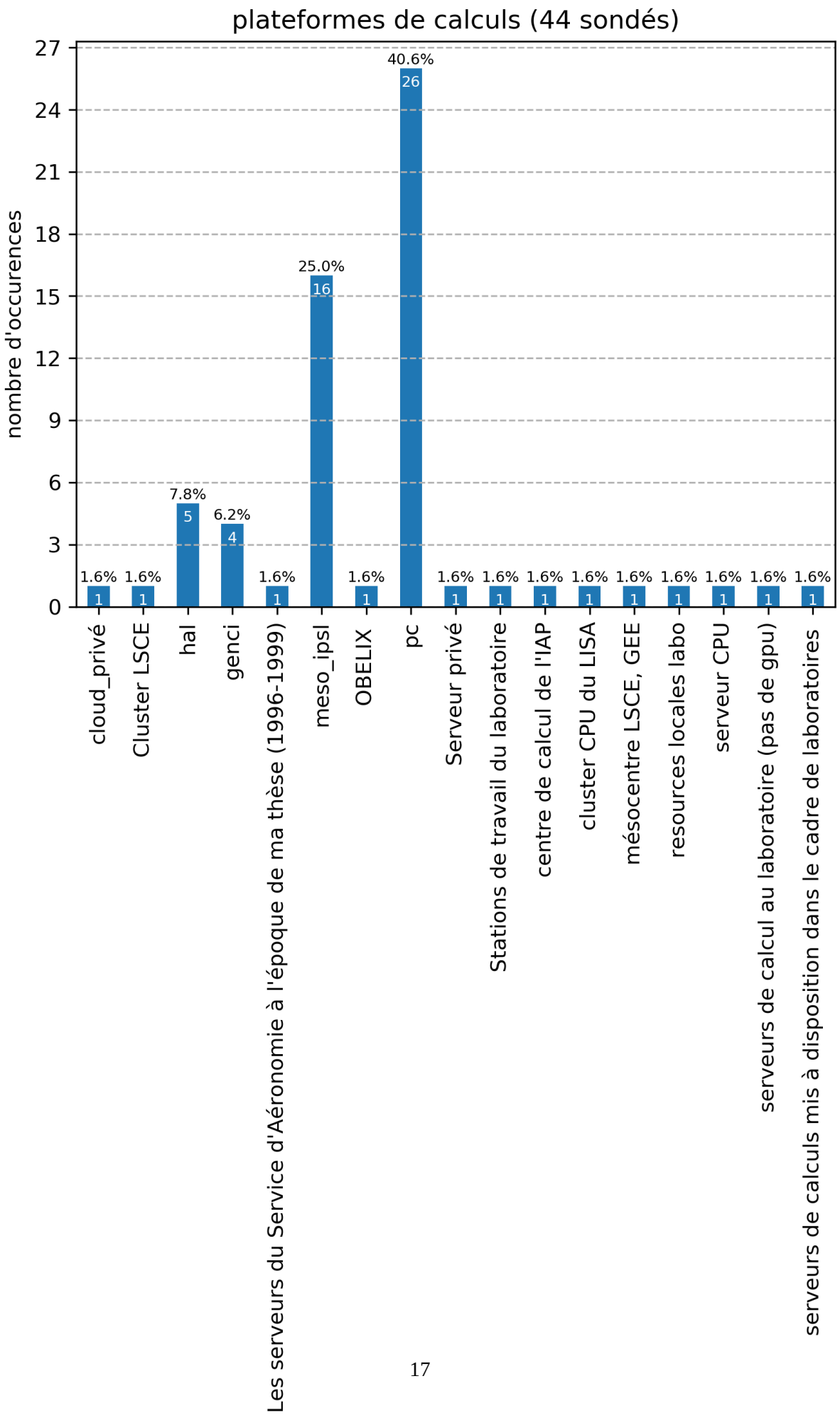
Figure 12 : projection ACP des usages ML × méthodes ML

2.9 Plateformes de calculs

L'utilisation des différentes plateformes de calculs pour les projets ML de 44 sondés est illustrée à la figure 13 (multiple). La machine personnelle (PC) est à l'écrasante majorité la plateforme de calculs des projets ML (40,6 %). Le mésocentre de l'IPSL (25 %), le cluster de GPU HAL (7,8 %) et les clusters de GENCI (6,2 %) sont bien loin derrière. On note un unique cas d'utilisation de cluster privé (GAFAM). Le détail de cette répartition par laboratoire est donné en annexe 30 (voir avertissement section 2.1).

L'analyse de la relation entre plateformes de calculs et méthodes ML (toutes les deux de type multiple) nous apprend qu'il existe trois groupes de sondés (voir *kmeans_centres_methodes_plateformes.csv*) :

- 13 sondés utilisant des méthodes non neuronales majoritairement sur le mésocentre de l'IPSL et sur leur PC de façon moins marquée que l'ensemble — le groupe du mésocentre —
- 9 sondés utilisant toutes les méthodes et toutes les plateformes de calculs.


plateformes de calculs (44 sondés)
nombre d'occurences
0
3
6
9
12
15
18
21
24
27
1.6%
1
cloud_privé
1.6%
1
Cluster LSCE
7.8%
5
hal
6.2%
4
genci
1.6%
1
Les serveurs du Service d'Aéronomie à l'époque de ma thèse (1996-1999)
25.0%
16
meso_ipsl
1.6%
1
OBELIX
40.6%
26
pc
1.6%
1
Serveur privé
1.6%
1
Stations de travail du laboratoire
1.6%
1
centre de calcul de l'IAP
1.6%
1
cluster CPU du LISA
1.6%
1
mésocentre LSCE, GEE
1.6%
1
resources locales labo
1.6%
1
serveur CPU
1.6%
1
serveurs de calcul au laboratoire (pas de gpu)
1.6%
1
serveurs de calculs mis à disposition dans le cadre de laboratoires

C’est le seul groupe à utiliser HAL — le groupe HAL —
- 19 sondés utilisant principalement MLP et les méthodes de régressions polynomiales principalement sur leur PC — le groupe PC —

Les trois groupes sont illustrés à la figure 14 (ACP). Les dendrogrammes de contrôle du nombre de groupes sont donnés en annexe 31 et 32.

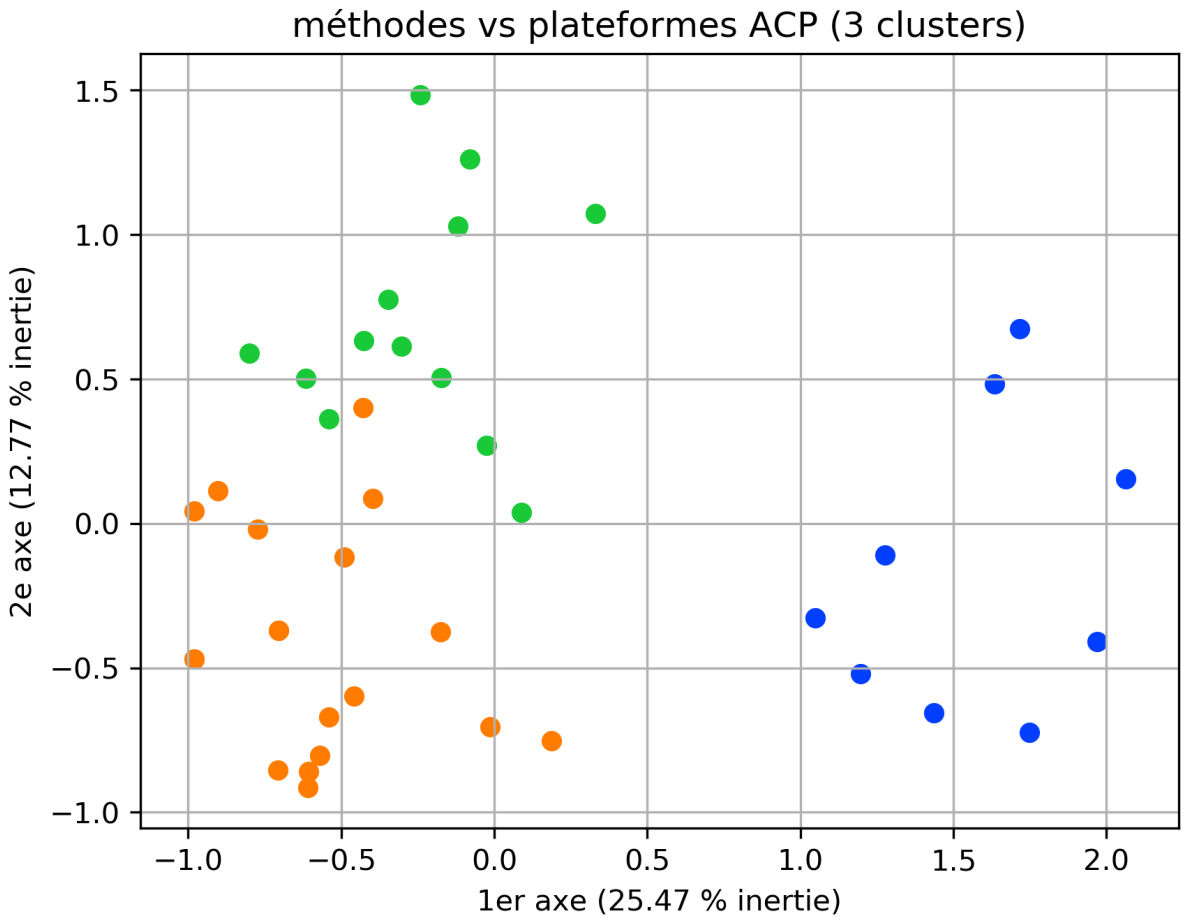


Figure 14 : projection ACP des plateformes de calculs × méthodes ML

2.10 Usage du GPU

La figure 15 donne un aperçu de l’usage des plateformes de calculs spécifiquement à base de GPU pour 42 sondés (multiple). On constate que les sondés utilisant des plateformes GPU représentent en tout 21,5 %, toutes d’architecture Nvidia. 7,1 % des sondés auraient pu utiliser des GPU pour leurs projets ML mais ne l’ont pas fait.

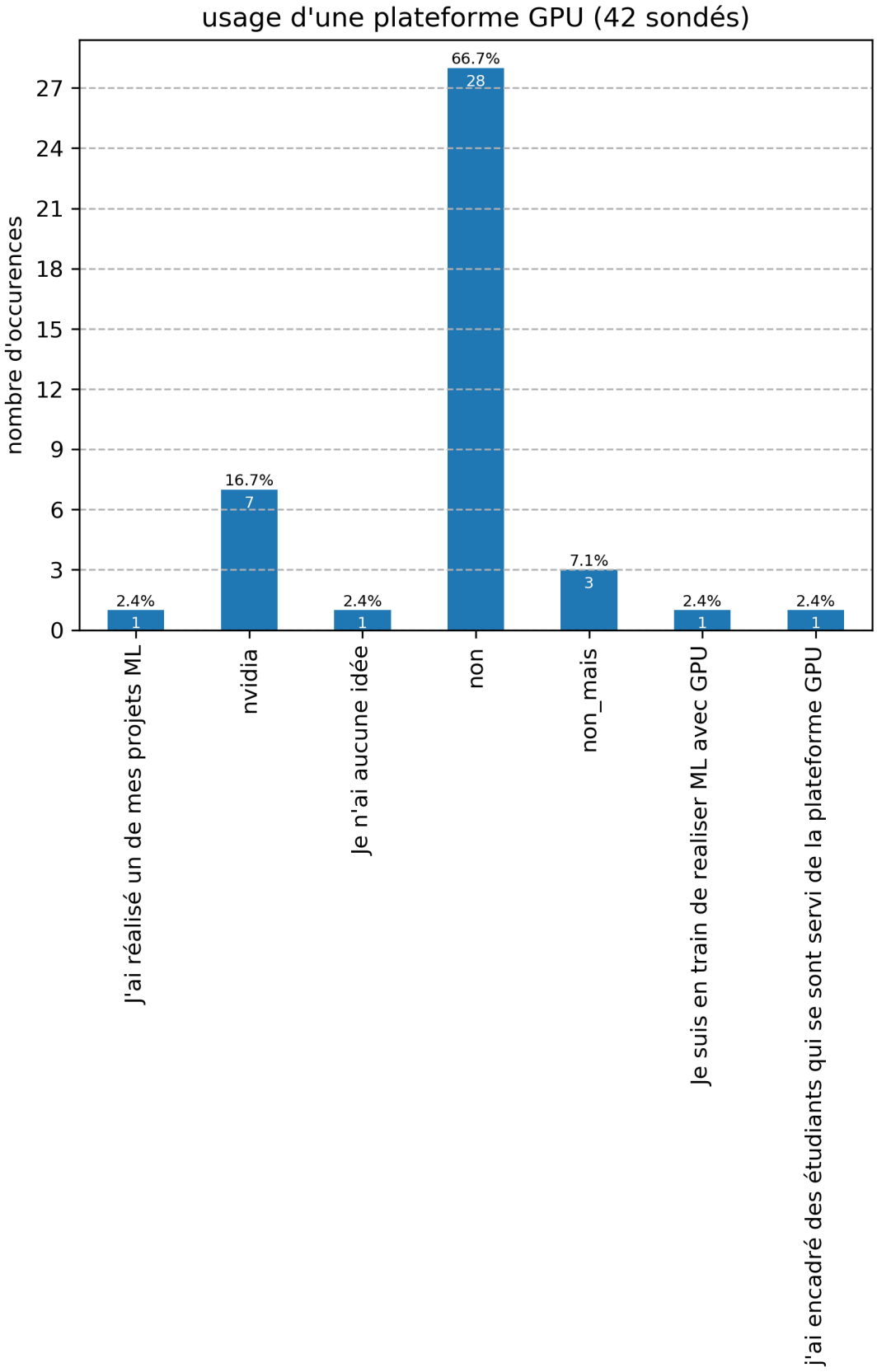


Figure 15 : répartition de l'utilisation de plateformes de calculs à base de GPU

2.11 Implémentation et outillage

Les réponses à la question de l'implémentation des projets ML, librement rédigées par les sondés, ont été interprétées afin d'en retirer pour chacune des mots clefs liés aux langages d'implémentation et à leurs bibliothèques. Donc langages et bibliothèques d'implémentation ont pu être analysés séparément.

2.11.1 Langages et bibliothèques

La figure 16 donne la répartition des langages d'implémentation utilisés par 42 sondés pour réaliser leurs projets ML. Sans surprise, Python (toutes versions confondues) est le langage de légèrement plus de la moitié des projets ML (55,3 %). Matlab et le langage R font respectivement 12,5 % et 8,9 %. Ces résultats sont cohérents avec les tendances actuelles en Data Science. À noter que les langages à base de JVM (Java, Scala, Kotlin, etc.) et le langage Julia (orienté Data Science) sont absents des langages cités par les sondés. La répartition par laboratoire donnée en annexe 33, ne dément pas la domination de Python au sein de tous les laboratoires.

La figure 17 montre l'utilisation des bibliothèques d'implémentation des projets ML de 22 sondés. Pour l'implémentation de projets DL, Keras (33,3 % ; Python) est devant Pytorch (15,2 % ; Python) et Tensorflow (12,1 % ; C++ ; Java ; Python). À noter que Keras est une surcouche de Tensorflow de plus haut niveau d'abstraction (plus simple d'utilisation) et est livrée avec Tensorflow depuis la version 2. Pytorch propose également une surcouche appelée Caffe mais elle n'est pas utilisée par les 22 sondés. Côté ML non neuronal, Scikit-learn (Python) domine avec 15,2 %.

2.11.2 Outillage

La figure 18 donne l'outillage utilisé par 43 sondés pour implémentation leurs projets ML (multiple). Sans surprise, Jupyter est l'outil le plus utilisé avec 27,5 % et le seul outil de type notebook apparaissant dans les usages des sondés. Par contre, on note que l'éditeur de texte (Emacs, Vim, etc.) est le deuxième outil le plus utilisé avec 22,5 %, score proche de celui de Jupyter. À noter que Jupyter n'est pas cantonné au langage Python, à l'aide d'extensions, il prend également en charge les langages C et Fortran. Matlab, langage mais également environnement intégré de développement (IDE), arrive en troisième avec 16,2 %. Concernant les IDE du monde Python, Spyder est préféré à Pycharm (respectivement 13,8 et 6,2 %). Spyder et Pycharm présentent tous les deux des facilités pour réaliser des projets ML (notebook, visualisation de données et de graphiques, etc.). La version communautaire de Pycharm, sous licence Apache, est une version

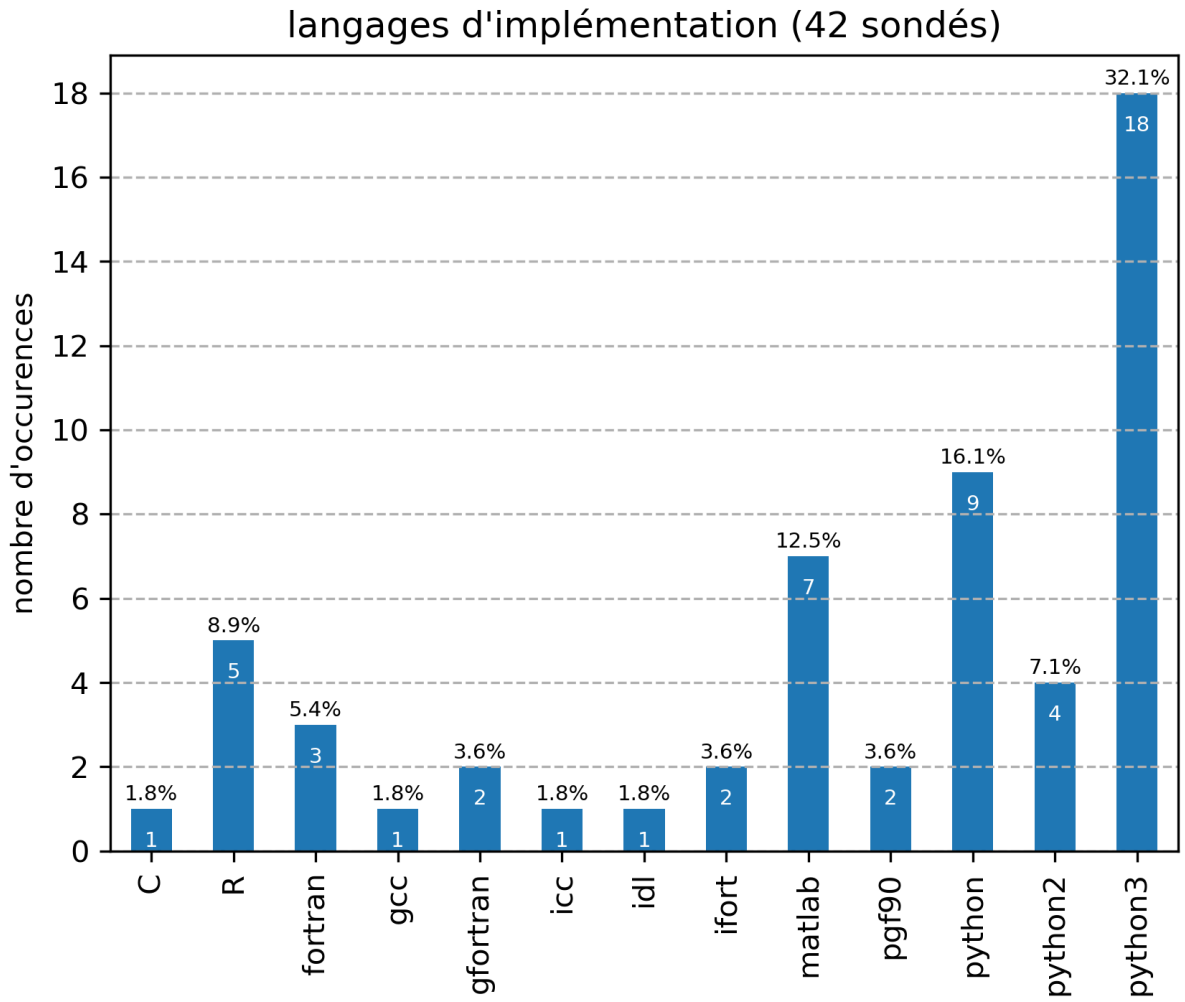


Figure 16 : utilisation des langages d'implémentation

avec moins de fonctionnalités que la version « professionnelle » de Pycharm. Si cette perte de fonctionnalité empêche l'adoption de Pycharm, il faut rappeler que la licence éducation, pleinement fonctionnelle, est gratuite pour les agents de la Sorbonne. L'analyse des données entre compétences et outillage n'a pas démontré statistiquement de corrélation (test du χ^2 négatif).

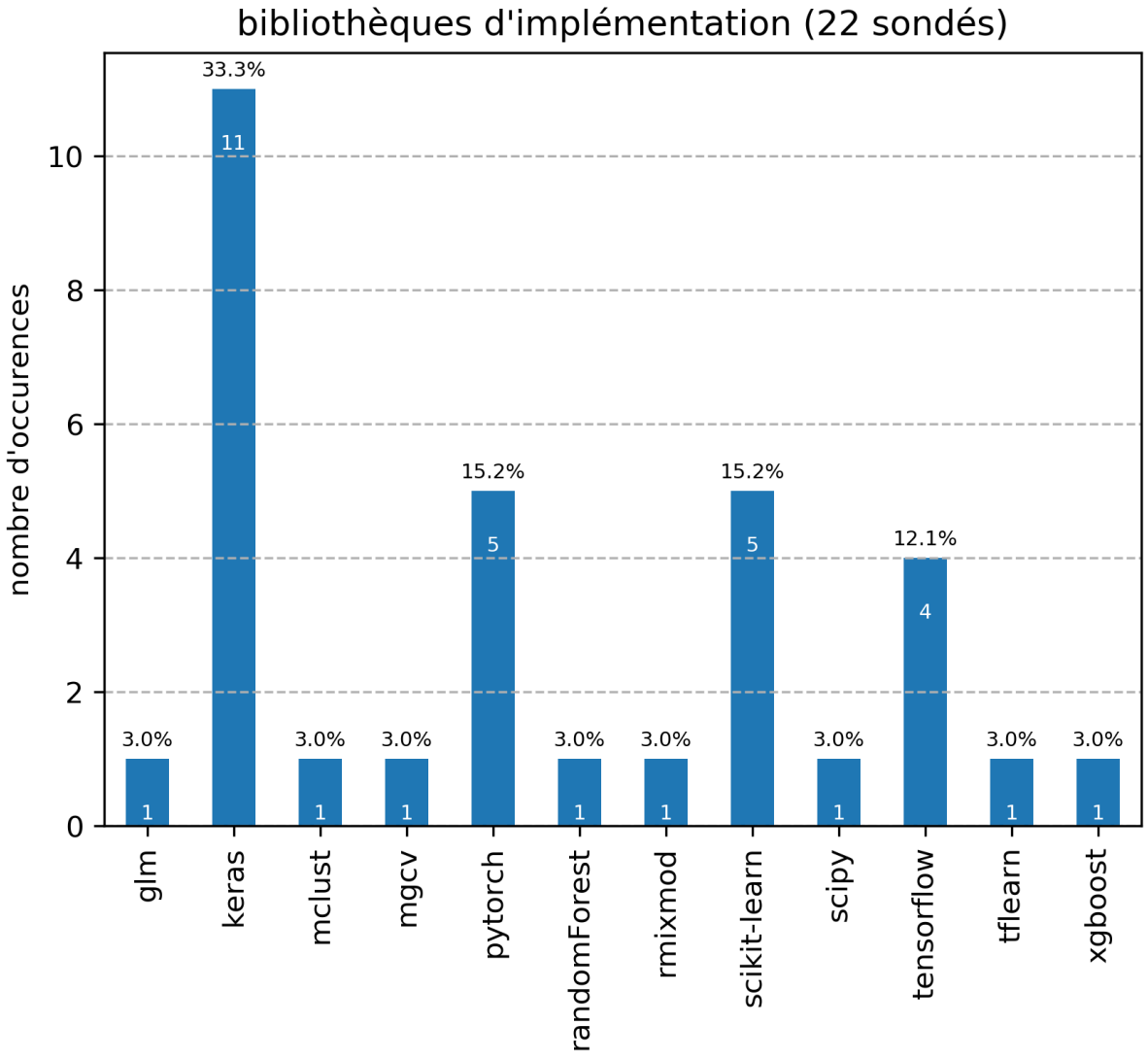


Figure 17 : utilisation de bibliothèques d'implémentation

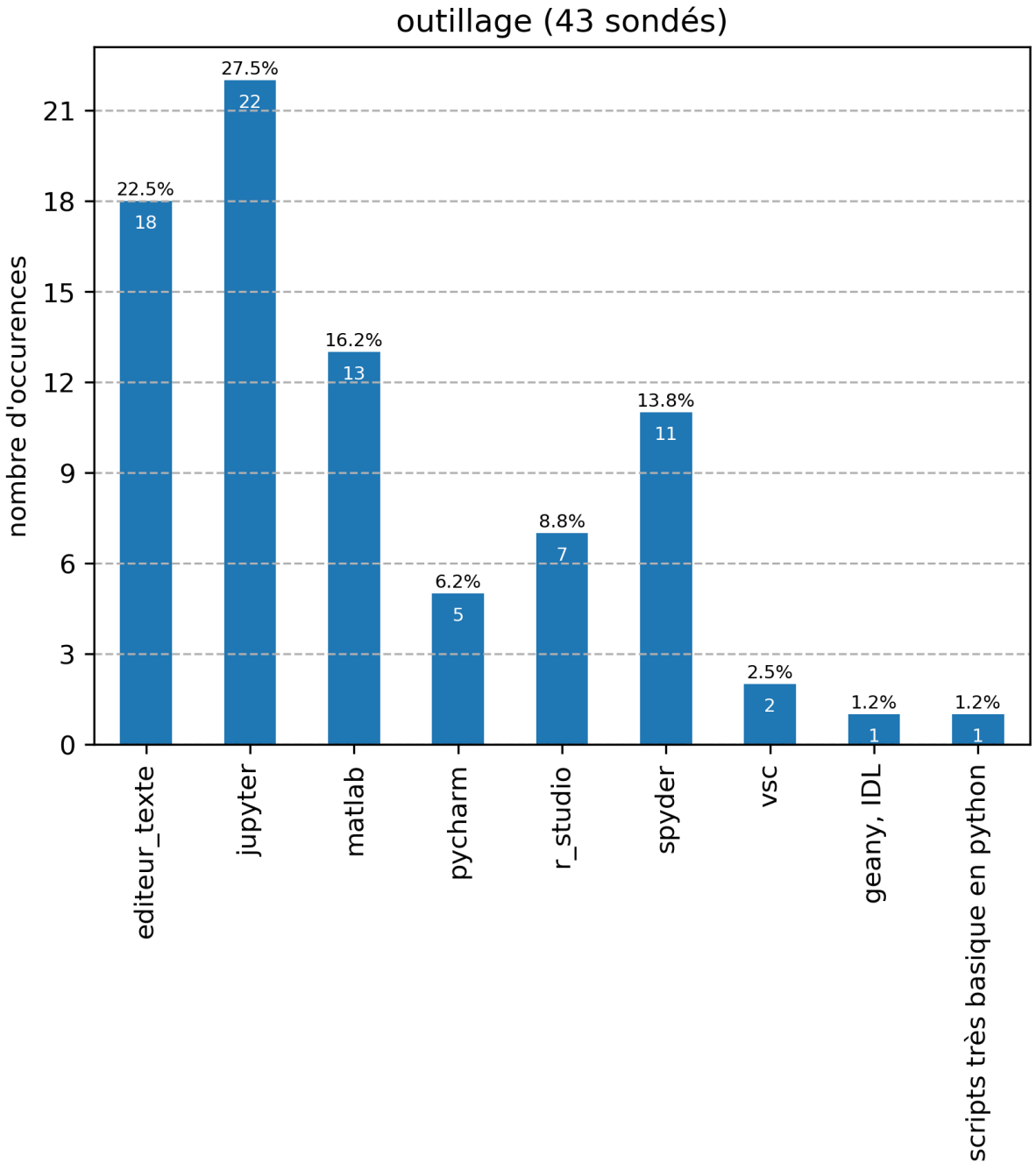


Figure 18 : répartition de l'outillage des projets ML

2.12 Big Data

L'avant dernière question a pour objet de déterminer si les sondés ont une problématique Big Data (multiple). Pour rappel, selon Laney 2001, le Big Data est traditionnellement présenté comme un ensemble de méthodes et technologies d'exploitation de très grands volumes de données ou de flux rapides de données, éventuellement non structurées (il n'existe pas de schéma de métadonnées en commun) : les fameux trois V pour Volume, Velocity et Variety. Fox 2018 complète cette définition en précisant que le traitement des Big Data nécessite des outils parallélisant les calculs, parallélisation entraînant la perte de certaines garanties ou capacités par rapport aux outils traditionnels implémentant le modèle relationnel (Codd 1970). Enfin Magoulas et Lorica 2009 indiquent que la notion de taille des Big Data dépend de l'infrastructure matérielle des organisations (pour certaines il s'agit de gigaoctets, pour d'autres de pétaoctets ou plus).

La figure 19 montre la répartition des problèmes posés par les trois V à 24 sondés. Arrive en tête le problème du volume de données (la question précisait supérieur à 10 To) avec 44,4 %, soit 12 personnes. Ensuite vient la variabilité des métadonnées (également spécifiées dans la question) avec 29,6 %, soit 8 personnes. Selon Dedić et Stanier 2017, le Big Data est principalement dédié aux données non structurées. S'il ne faut garder que les sondés ayant déclaré avoir des problèmes de volume et de variabilité ou de vélocité et de variabilité, il ne reste plus qu'une seule personne. D'autre part, il y a une différence entre traiter un jeu de données Big Data et traiter un sous ensemble de ce jeu de données que l'on ne peut pas qualifier de Big Data. Cette différence n'est malheureusement pas explicitée dans la question, pas plus que la nécessité de paralléliser et distribuer les traitements. Je pense qu'il est difficile de conclure si les sondés sont réellement aux prises avec des problématiques Big Data. Un complément est donc nécessaire afin de mieux comprendre les besoins en termes de Big Data dans les laboratoires de l'IPSL.

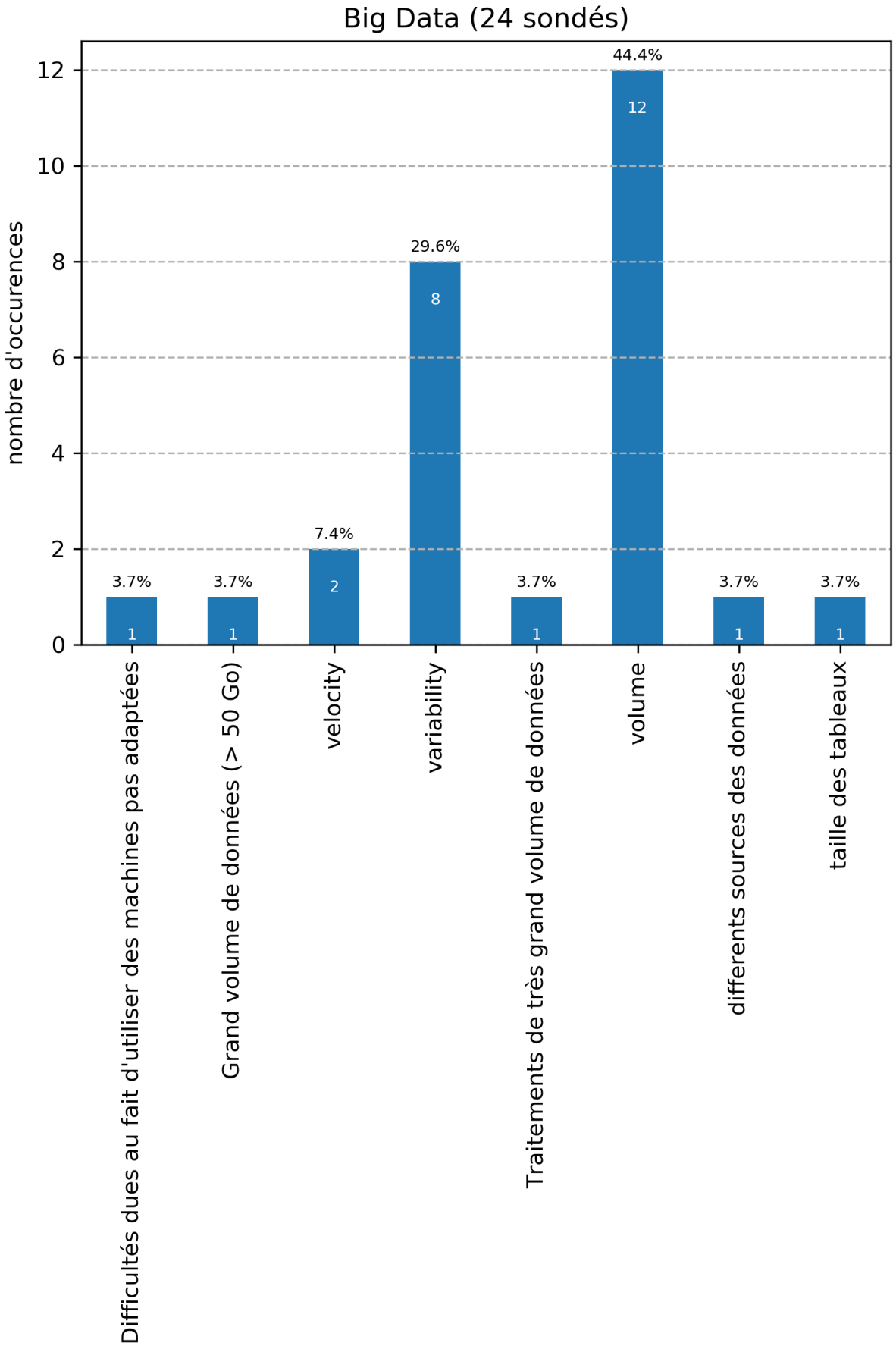


Figure 19 : répartition des sondés confrontés aux trois V du Big Data

2.13 Attentes ESPRI-IA

La figure 20 montre la distribution des attentes de 37 sondés concernant ESPRI-IA, basées sur une interprétation de leurs réponses sous forme de mots clefs. Bien sûr, il n'est pas possible de retranscrire en mots clefs toutes les nuances des réponses, aussi il est encouragé de lire les réponses in extenso dans le fichier *extended_sondage.csv*. Majoritairement, les sondés attendent d'ESPRI-IA des conseils pour la réalisation de leurs projets IA. Des conseils à propos des méthodes ML, d'ingénierie, d'implémentation, sur les infrastructures matérielles à l'IPSL ou à l'extérieur. Vient ensuite la création d'une communauté autour de l'IA qui favoriserait l'entre-aide, le partage d'expériences, la rencontre entre pratiquants de l'IA afin de favoriser des échanges voir des actions communes. En troisième position, on trouve le besoin de mutualiser les ressources (matérielles et logicielles), et l'accès aux moyens de calculs. La formation est également une demande relativement forte. Pour information, le mot clef « accompagnement » désigne une aide plus poussée que le conseil, il regroupe les sondés souhaitant une contribution active d'ESPRI-IA pour la réalisation de leurs projets. Le niveau ultime étant représenté par le mot clef « implémentation » qui désigne les personnes souhaitant qu'ESPRI-IA prenne intégralement en charge leurs projets. Le mot clef « référencement projet IA », est le souhait d'organiser la centralisation des livrables de projets IA et leur indexation par mots clefs. Le mot clef « séminaires » regroupent les souhaits de suivre des séminaires d'applications scientifiques de l'IA dans le but d'évaluer l'utilité de ses méthodes. Enfin, « dev bibliothèques » rassemble les sondés souhaitant le développement de bibliothèques ML réutilisables pour des projets scientifiques.

3 Bilan

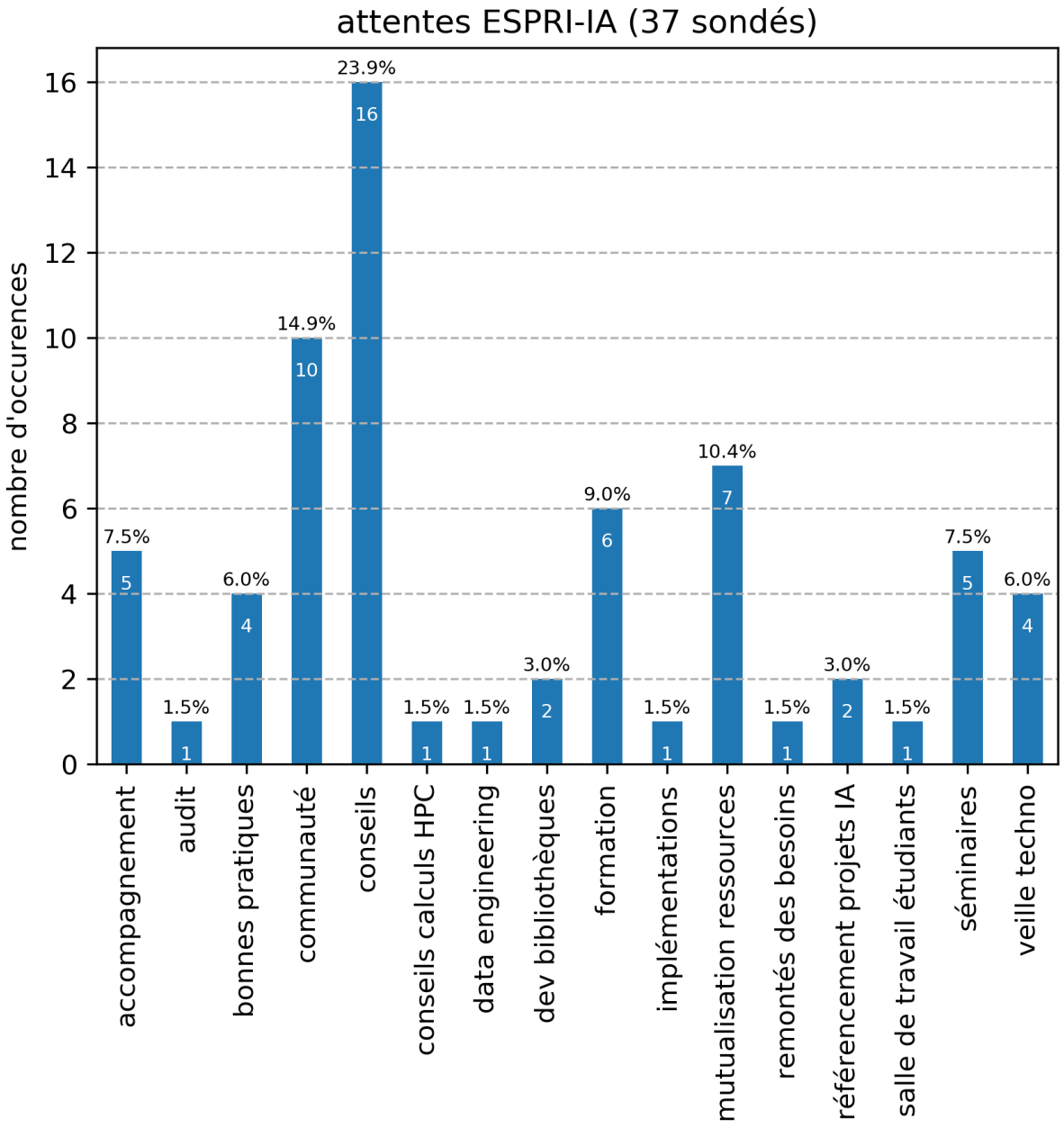


Figure 20 : interprétation des attentes concernant ESPRI-IA

3.1 Synthèses des analyses

Le sondage a connu une bonne participation avec 62 sondés. Cependant, il faut tenir compte dans les analyses de la sur-représentation des membres du LATMOS, LMD et LSCE, surtout pour les analyses bidimensionnelles basées sur le rattachement du sondé à un laboratoire. On constate que la moitié des sondés n'ont pas de compétences en Machine Learning (ML) mais sont intéressés par ses méthodes. Le sondage révèle qu'au moins trente sondés ont besoin d'aide pour réaliser leurs projets ou idées de projets ML. Les thèmes de la simulation climatique, de l'atmosphère et de la télédétection regroupent la majorité des sondés intéressés par le ML. La régression, classification et le clustering sont les principaux usages du ML dans les laboratoires de l'IPSL. Les méthodes ML non neuronales sont en tête des pratiques, les méthodes neuronales représentent 27,3 %. Cependant, il faut tenir compte d'un biais entre la compétence des sondés et leur pratique du ML. Le ML se pratique à l'écrasante majorité sur une machine personnelle sur des données provenant de CMIP5&6, de ERA5, de SIRTA, de la NASA et de IASI. L'implémentation des projets ML est dominée par l'écosystème formé par Python et ses bibliothèques phares (Keras, Tensorflow, Pytorch et Scikit-learn). Jupyter Notebook et les éditeurs de texte dominent l'outillage. Enfin, les sondés attendent majoritairement d'ESPRI-IA des conseils pour la réalisation de leurs projets ML, la formation d'une communauté autour de l'IA à l'IPSL et la mutualisation des ressources matérielles et logicielles.

3.2 Prescriptions

Suite à l'analyse du sondage, l'auteur de ce document propose une liste de sujets à aborder lors d'une réunion ESPRI-IA. Dans un ordre quelconque :

- L'organisation des Travaux Dirigés sous forme d'hackathon ML afin de mettre en pratique les méthodes ML sous Jupyter Notebook.
- L'expérimentation des extensions Fortran et C pour Jupyter Notebook.
- Le développement de l'animation autour du langage Python et/ou participer à l'initiative de formation Python du LATMOS.
- La formation systématique en ML (externe ou interne) de tout nouvel arrivant (stagiaires M2, doctorants et permanents).
- L'implémentation de la tâche de conseils d'ESPRI-IA.
- La réflexion sur la meilleure implémentation du concept de communauté IA.
- La mutualisation des ressources matérielles et de l'accès aux ressources de calculs.

- La mutualisation des ressources logicielles (en prenant l’exemple de Jean-Zay ?).

ANNEXES

A Compléments

A.1 Mots clefs

A.1.1 Questions/Noms de colonne

Horodateur → date

A quel laboratoire êtes vous rattaché·e ? → lab

Quels sont vos thèmes scientifiques et/ou vos domaines dingénierie, concernés par le Machine Learning ? → thème

Quel est votre degré de compétences en Machine Learning ? → compétence

Quel est le degré de maturité de vos projets Machine Learning ? → maturité_projet

Quelles sont les origines de vos données concernées par le Machine Learning ? → data_source

Avez vous des partenariats avec des laboratoires extérieurs à lIPSL ou des entreprises, concernant les aspects Machine Learning de vos projets ? → partenaires

Quelles sont vos attentes ou remarques au sujet dESPRI-IA ? → espri-ia_rqs

Si vous souhaitez être contacté·e, laissez nous votre email à cet endroit (optionnel) → email

Quelles méthodes ML avez vous déjà mises en œuvre (plusieurs choix possibles) ? → ml_méthodes

Quels objectifs avaient vos modèles ML (plusieurs choix possibles) ? → ml_usages

Quelles plateformes de calculs avez vous utilisées pour vos projet ML (plusieurs choix possibles) ? → exec_plateformes

Avez vous utilisé une plateforme GPU pour vos projets ML (plusieurs choix possibles) ? → gpu_projets

Quels sont les langages informatiques, bibliothèques et plateforme dexécution (CPU et/ou GPU) utilisés pour vos projets ML ? Si le langage est compilé (C, Fortran, etc.), veuillez indiquer le nom du compilateur. → implémentations

Quels outils de programmation avez vous utilisé pour réaliser vos projets ML (plusieurs choix possibles) ? → outillages

Avez vous eu des problèmes pour traiter vos données pour vos projets Machine Learning (plusieurs choix possibles) ? → big_data

Avez vous des commentaires ? → commentaires

A.1.2 Compétences

Je n'ai pas de compétence en ML mais je suis intéressé·e par le ML → interet
J'ai des compétences mais j'ai des difficultés pour réaliser des projets en ML → non_autonome
J'ai des compétences suffisantes pour réaliser des projets en ML → autonome

A.1.3 Maturité projets

Je n'ai pas de projet ML → aucun
Je n'ai pas de projet ML mais j'ai des données valorisables → données_valorisables
J'ai une idée ou un projet ML mais je ne l'ai pas encore réalisé → idées
J'ai réalisé ou fait réaliser des projets ML à l'aide de méthodes de ML classiques → classiques
J'ai réalisé ou fait réaliser des projets de ML qui ont nécessité la création de nouvelles méthodes (ou la modification de méthodes classiques) → custom

A.1.4 Méthodes ML

K-Nearest Neighbour, Kmeans et variantes → kmeans
Classification Ascendante/Descendante Hiérarchique (CAH et CDH) → cah/cdh
SOM (dites cartes auto-organisatrices ou cartes de Kohonen) → som
Méthodes factorielles (EOF, ACP, AFC, ACM, AFDM, AFM, etc.) → méthodes_fac
Perceptron Multi Couches (MLP) → mlp
Deep Learning (CNN, Auto-encodeurs, etc.) → dl
Machines à vecteurs de support (SVM) → svm
Random Forest et variantes → random_forest
Gradient Boosting et variantes → gradient_boosting
Algorithmes génétiques → algo_génétique
Régression linéaire ou polynomiale multivariée → régressions
Régression logistique → reg_log

A.1.5 Usages ML

Classification (supervisé) → classification
Détection → détection
Segmentation → segmentation
Prédiction → prédiction
Clustering (non supervisé) → clustering
Régression → régression
Réduction de dimension - Analyse d'espaces latents → réduction_dim

Filtrage/Débruitage de signaux → filtrage
Génération de données → génération
Imputation de données manquantes → imputation
Downscaling (extrapolation de résolution) → downscaling

A.1.6 Plateformes de calculs

Cluster de GPU HAL (OVSQ/LATMOS) → hal
Cloud privé (Amazon AWS, Microsoft Azure, Google Cloud, IBM Cloud, etc.) → cloud_privé
Genci (Curie, Irène, Jean Zay, etc.) → genci
Mésocentre IPSL (Ciclad & Climserv) → meso_ipsl
Ordinateur personnel → pc

A.1.7 Plateformes GPU

Je n'ai pas réalisé de projet ML sur plateforme GPU → non
Je n'ai pas réalisé de projet ML sur plateforme GPU mais j'aurai pu/dû → non_mais
J'ai réalisé un de mes projets ML sur plateforme GPU NVIDIA → nvidia
J'ai réalisé un de mes projets ML sur plateforme GPU AMD → amd

A.1.8 Outillage

Eclipse → eclipse
Editeurs de texte (Atom, Emacs, Sublime Texte, Vim, etc.) → editeur_texte
Jupyter Notebook → jupyter
Matlab → matlab
Netbeans → netbeans
Pycharm → pycharm
R studio → r_studio
Spyder → spyder
Visual Studio Code → vsc

A.1.9 Big Data

Traitements de très grand volume de données à faible densité d'information (> 10 To) → volume
Traitements de données à grande variabilité de métadonnées → variability
Traitements de données produites en continu très rapidement (streaming de données) → velocity

A.2 Dépendances

Les bibliothèques Python3.7 suivantes sont nécessaires afin d'exécuter le script *sondage.py* :

- Matplotlib 3.1.3
- Numpy 1.18.1
- Pandas 1.0.3
- Scipy 1.4.1
- Seaborn 0.10.1
- Sklearn 0.22.1

B Graphiques

B.1 Cartes de fréquentation

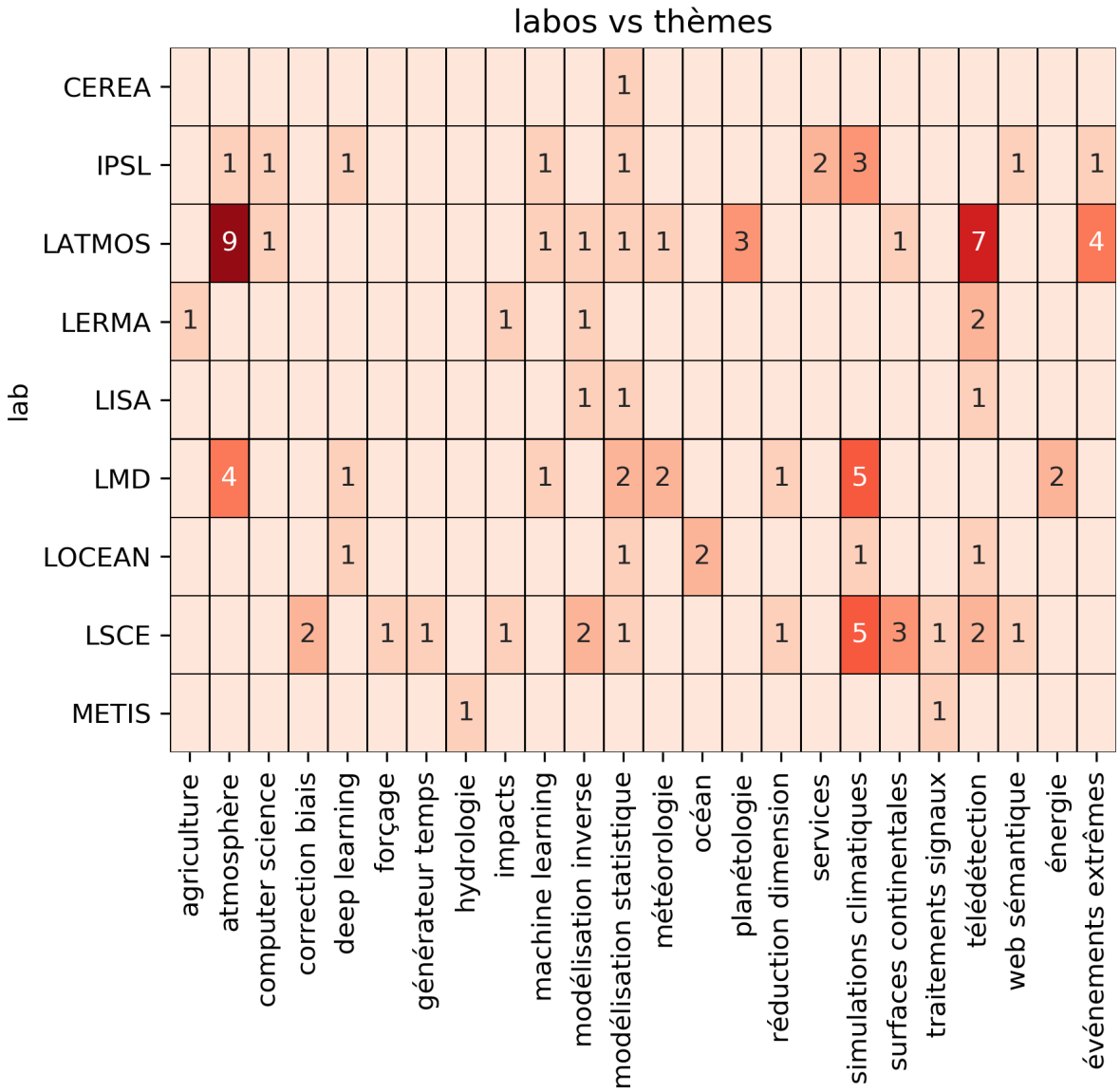


Figure 21 : thèmes scientifiques et techniques par laboratoire

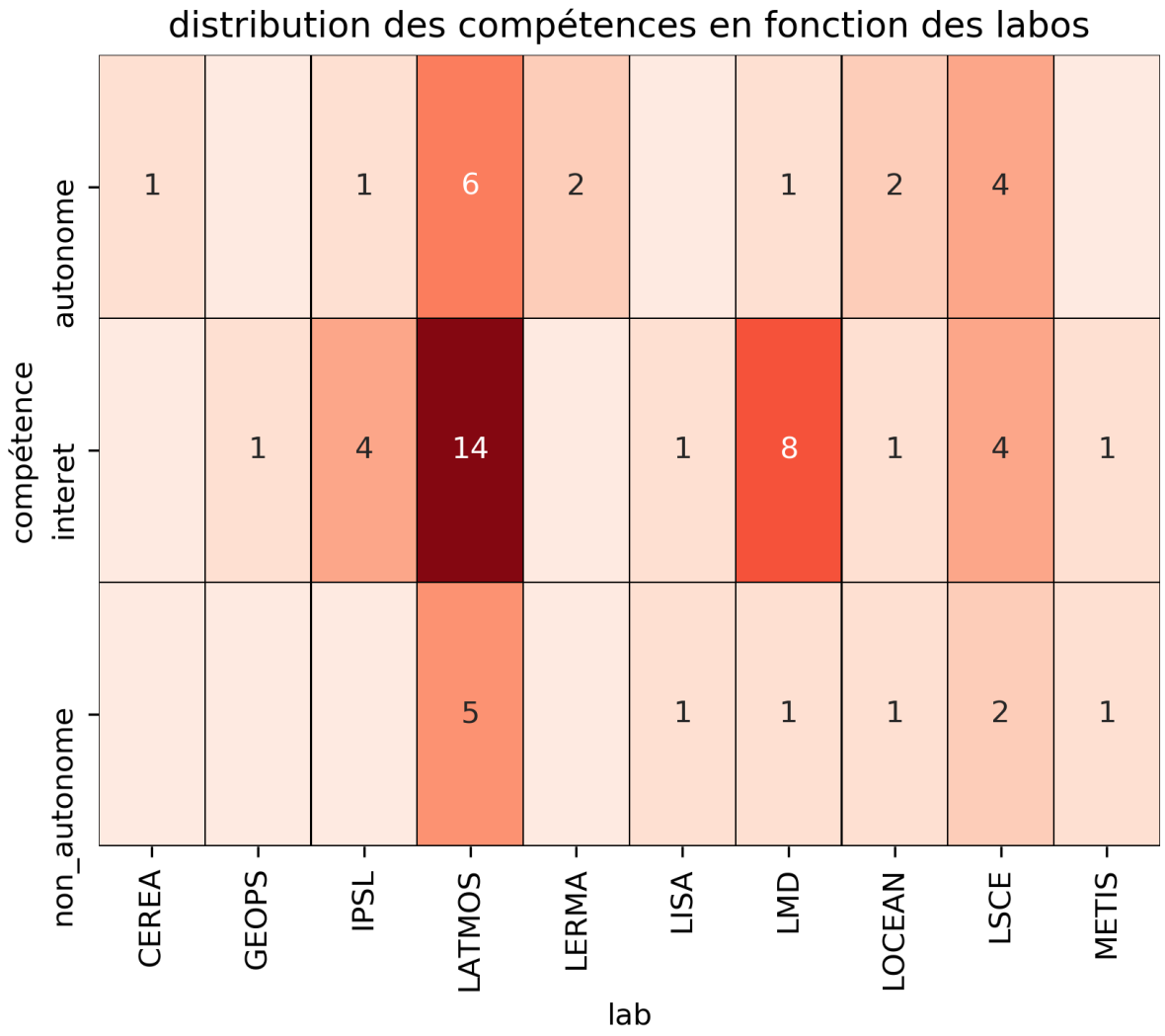


Figure 22 : compétences des sondés par laboratoire

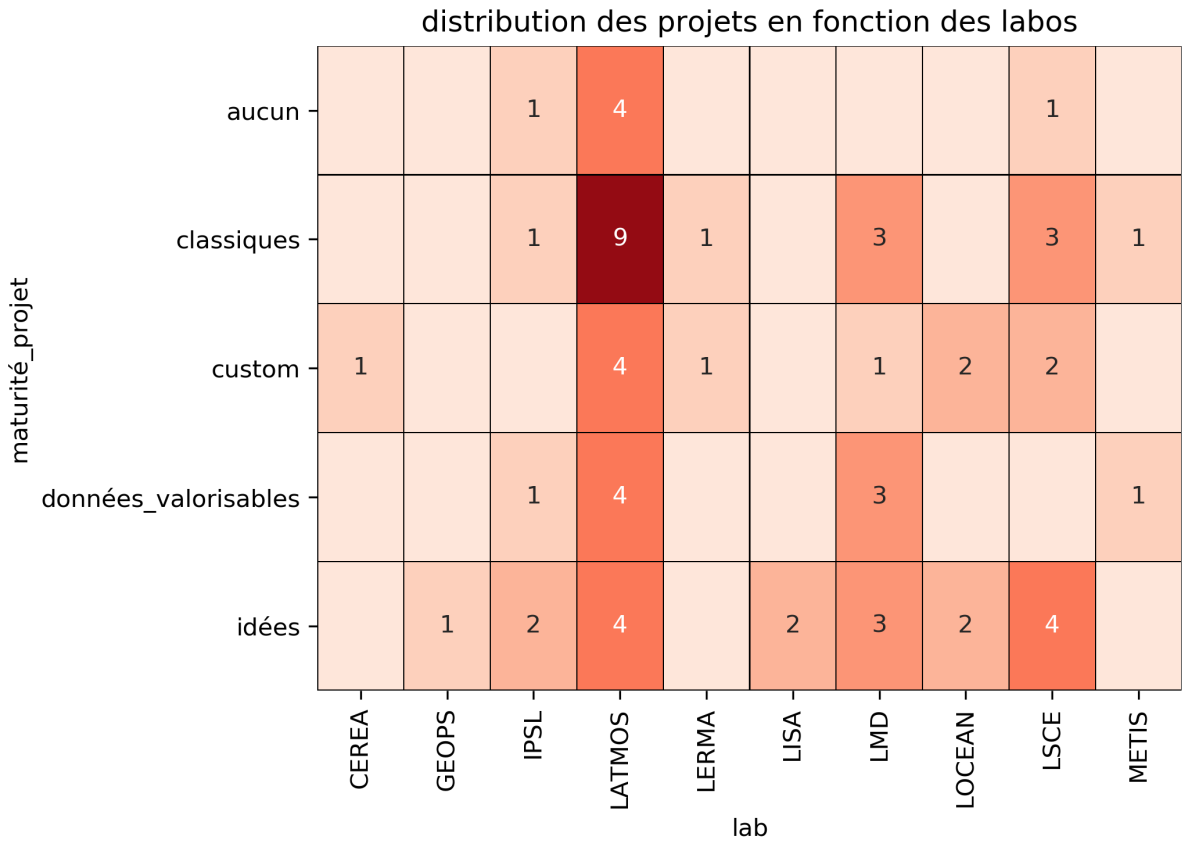


Figure 23 : degré de maturité des projets par laboratoire

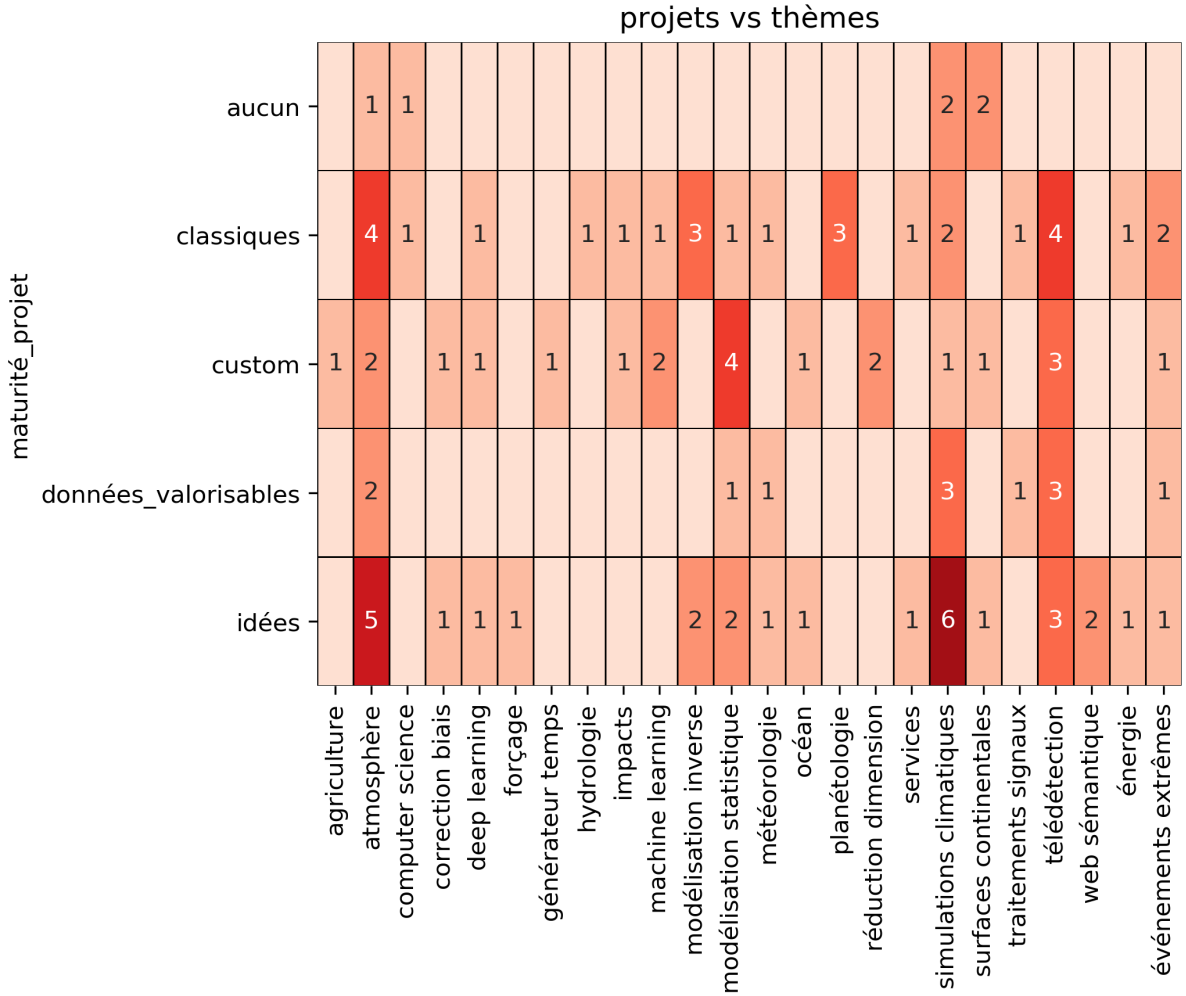


Figure 24 : degré de maturité des projets par thèmes scientifiques et techniques

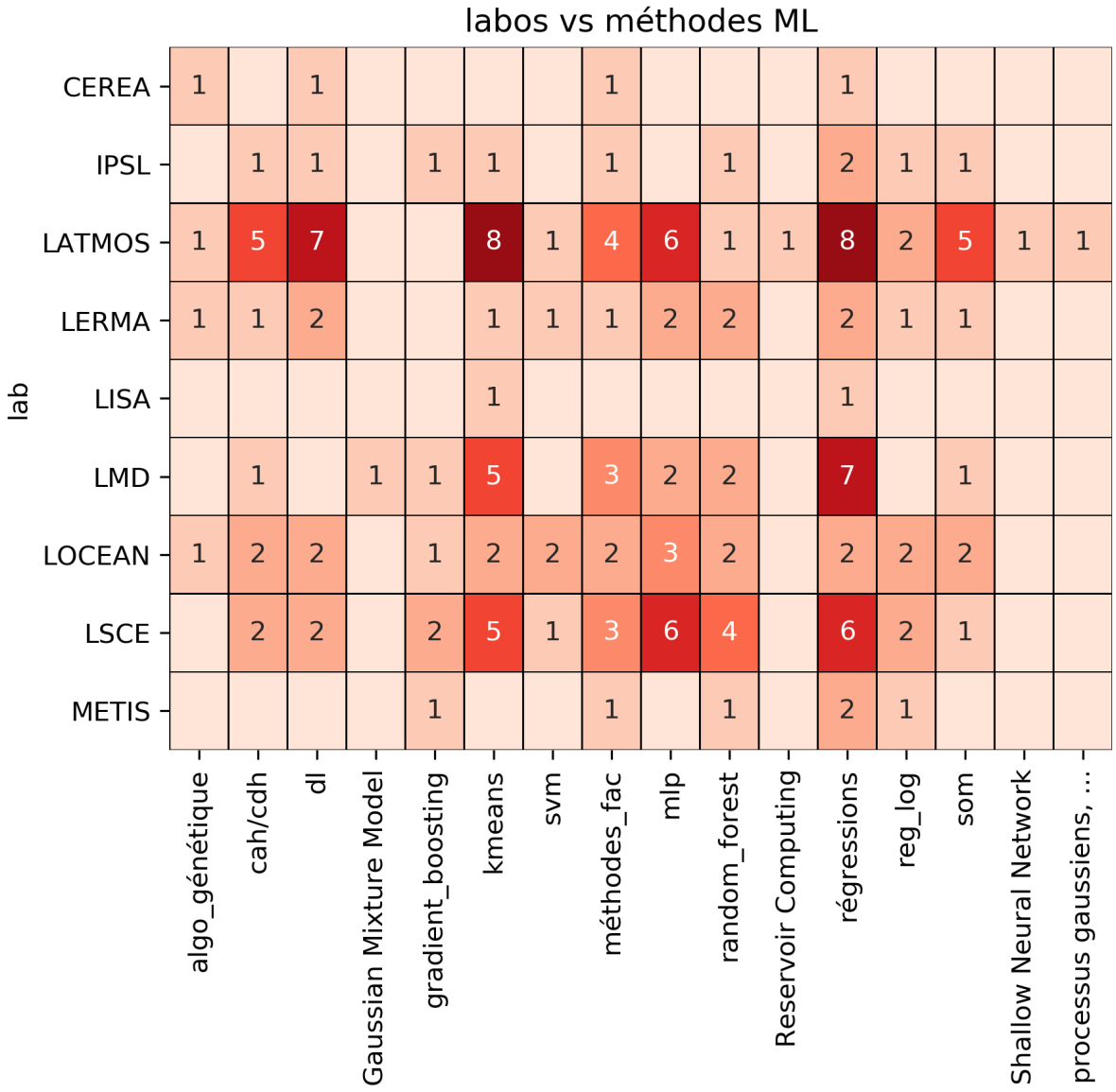


Figure 25 : utilisation des méthodes ML par laboratoire

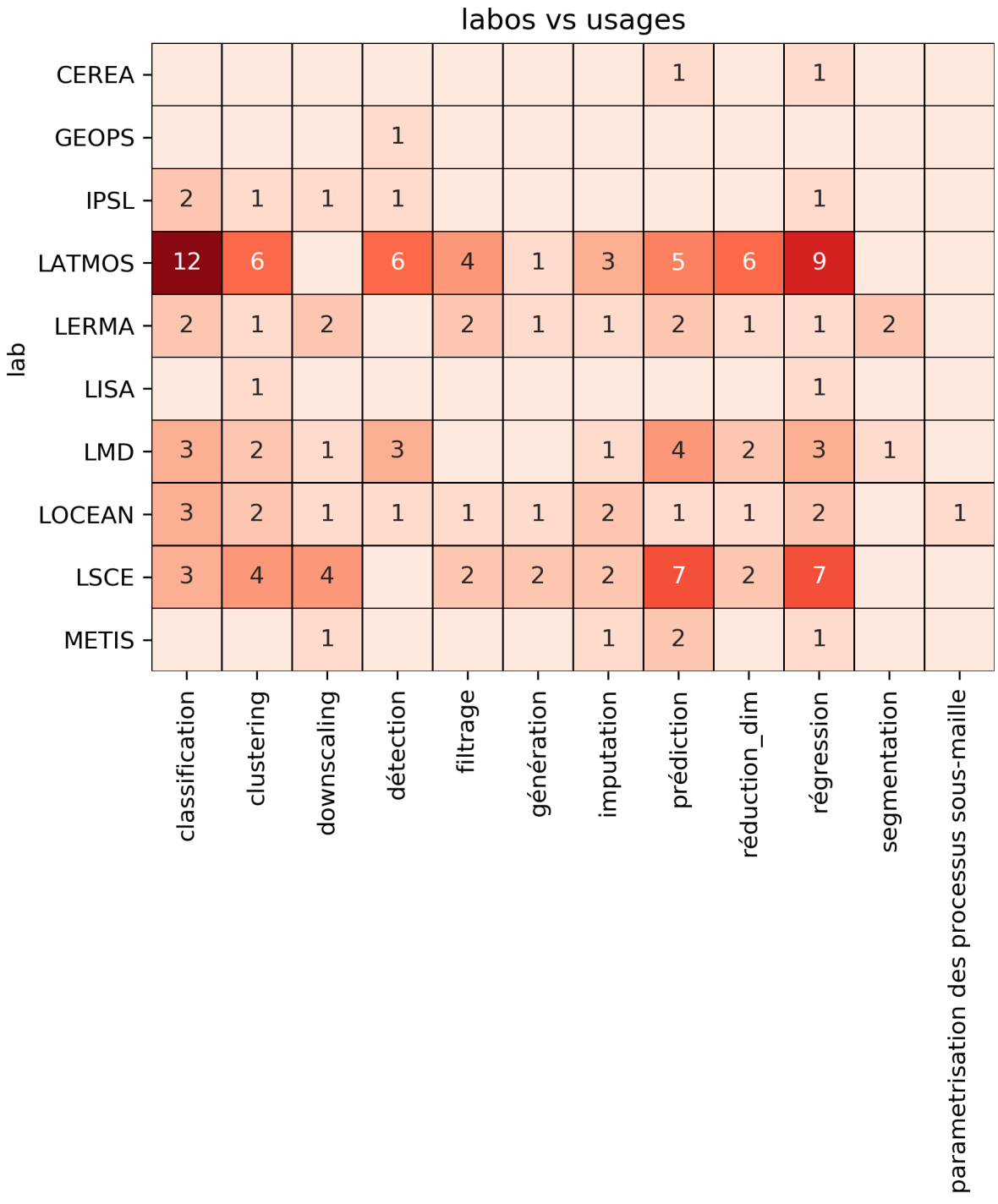


Figure 26 : usage du ML par laboratoire

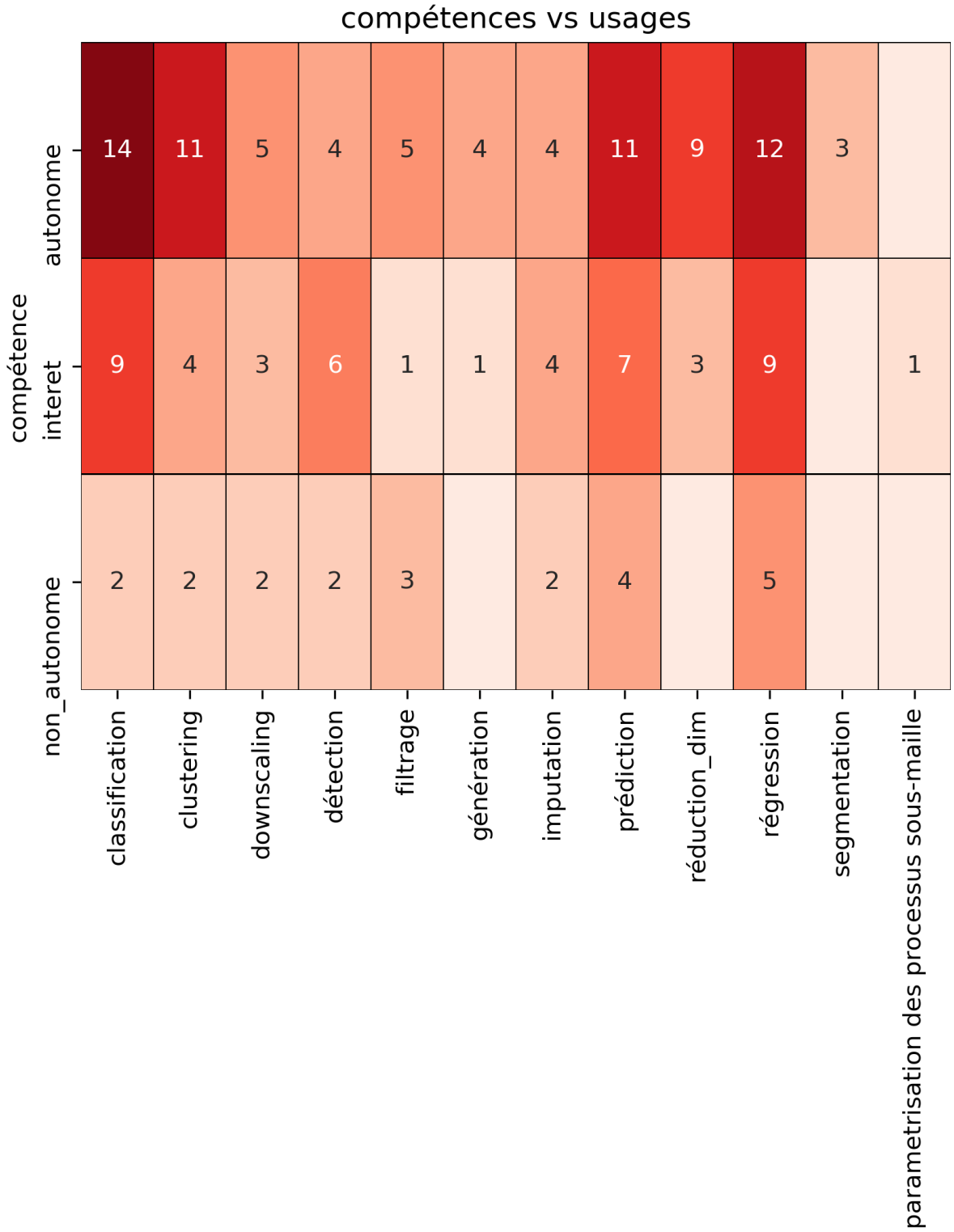


Figure 27 : usage du ML par compétences des sondés

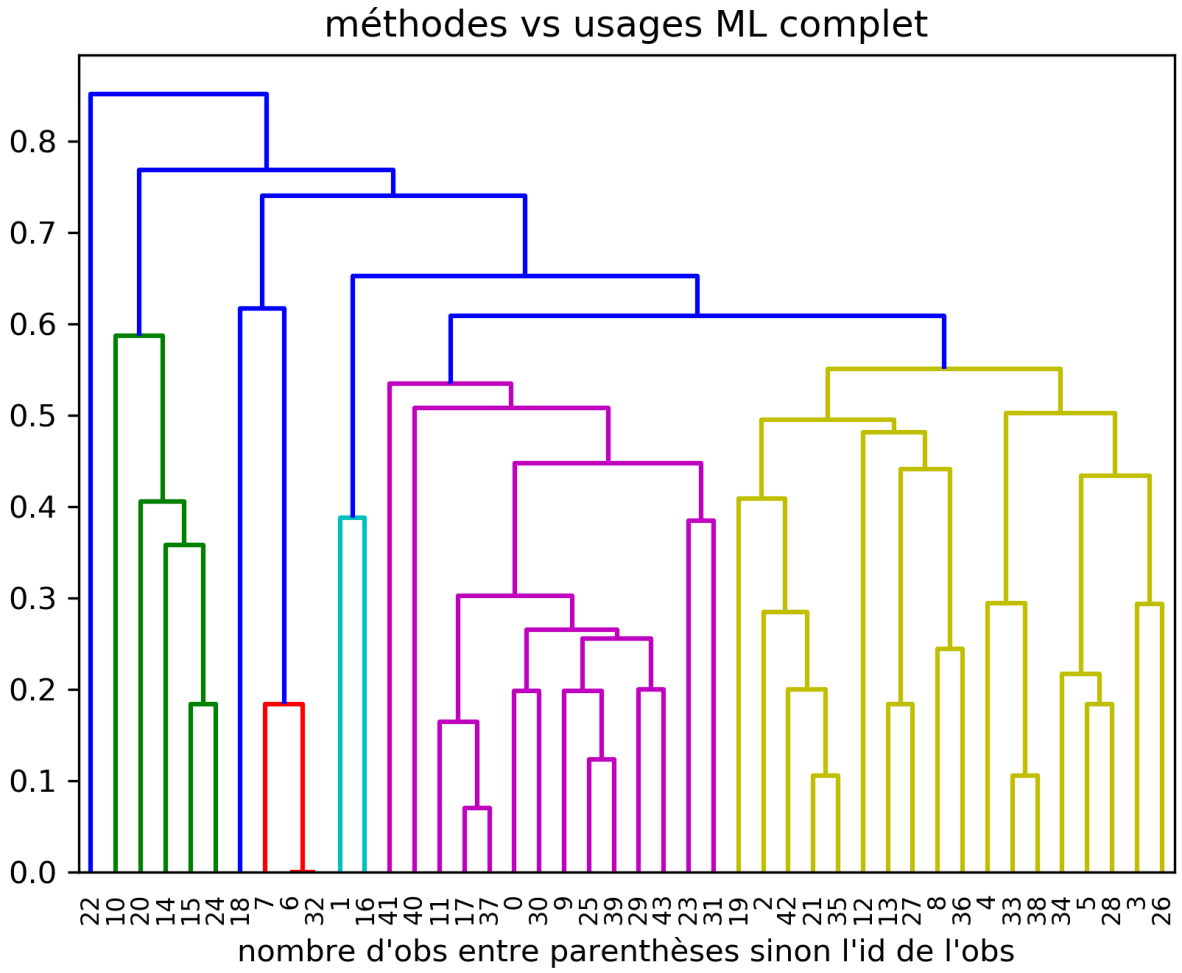


Figure 28 : dendrogramme du clustering méthodes et usages ML, arbre complet

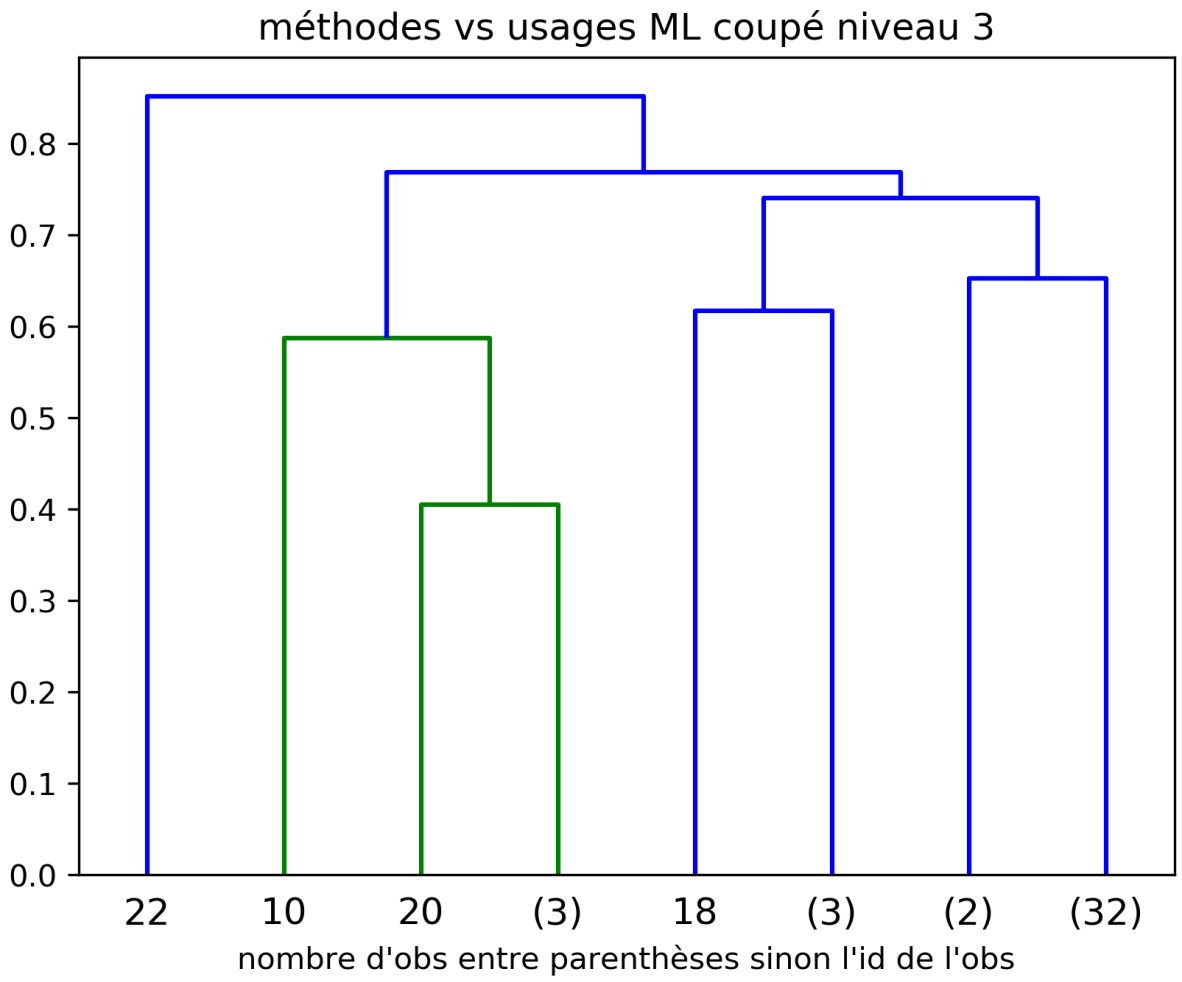


Figure 29 : dendrogramme du clustering méthodes et usages ML, coupé au niveau 3

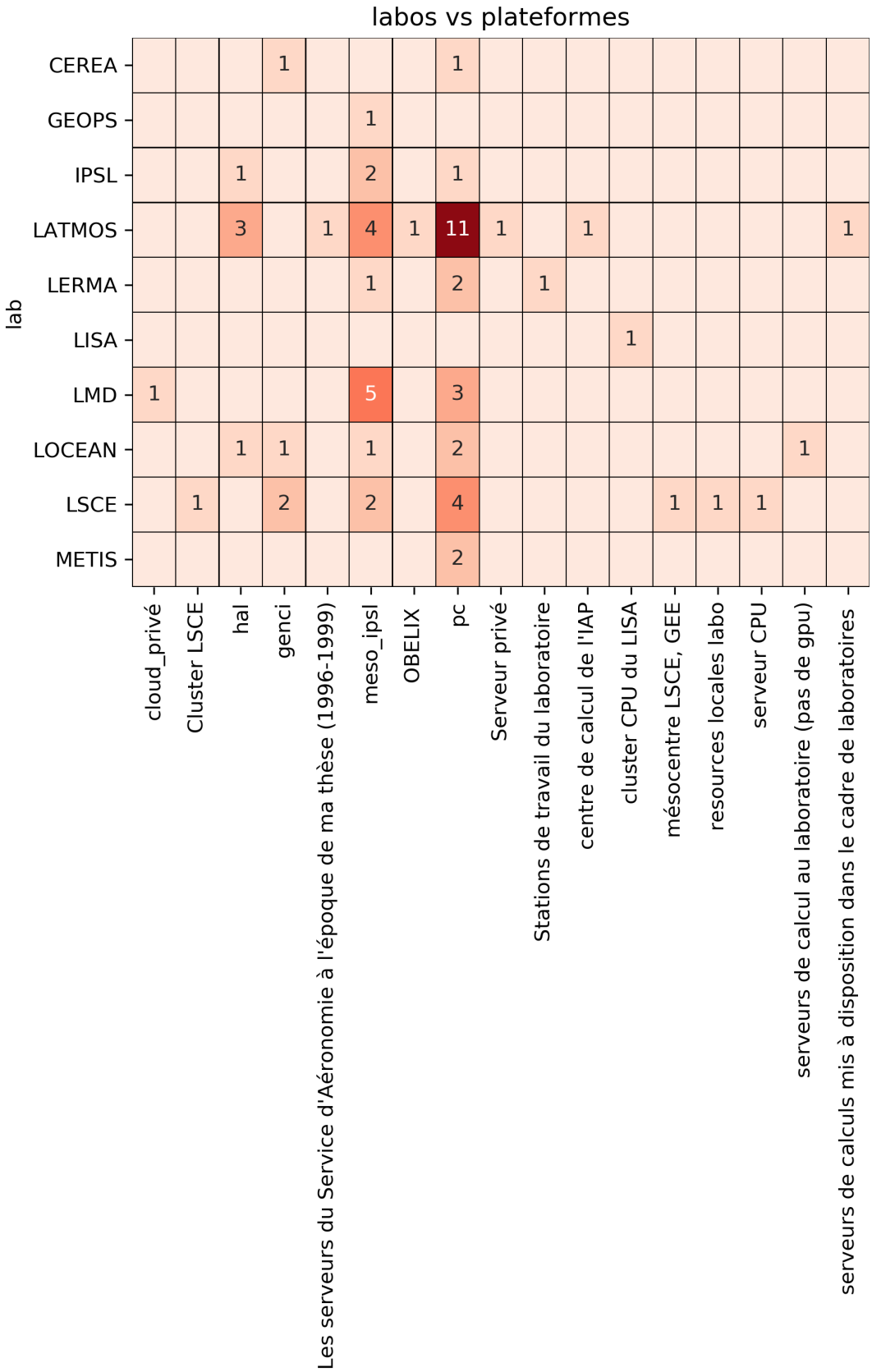


Figure 30 : utilisation des plateformes de calculs par laboratoire

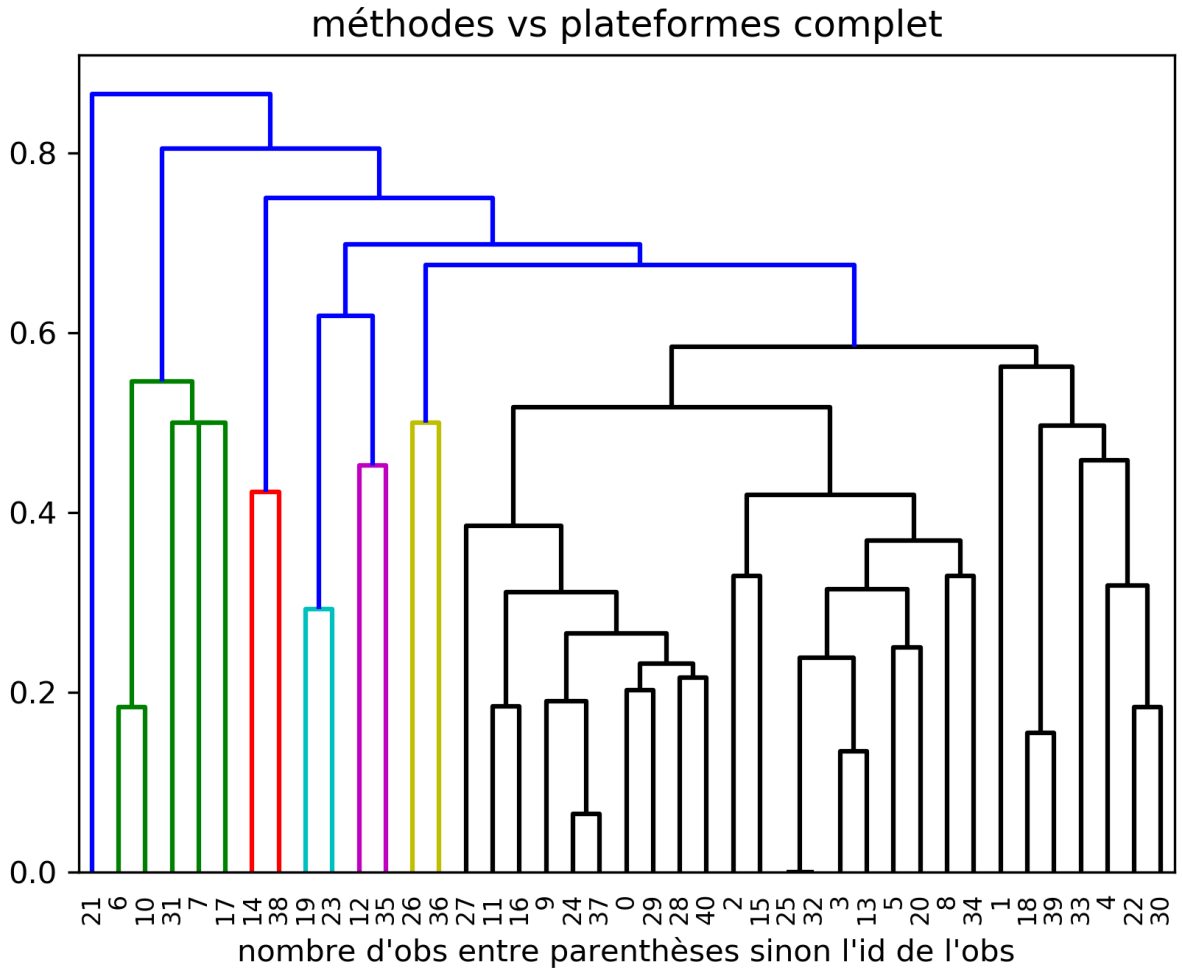


Figure 31 : dendrogramme du clustering méthodes ML et plateformes de calculs, arbre complet

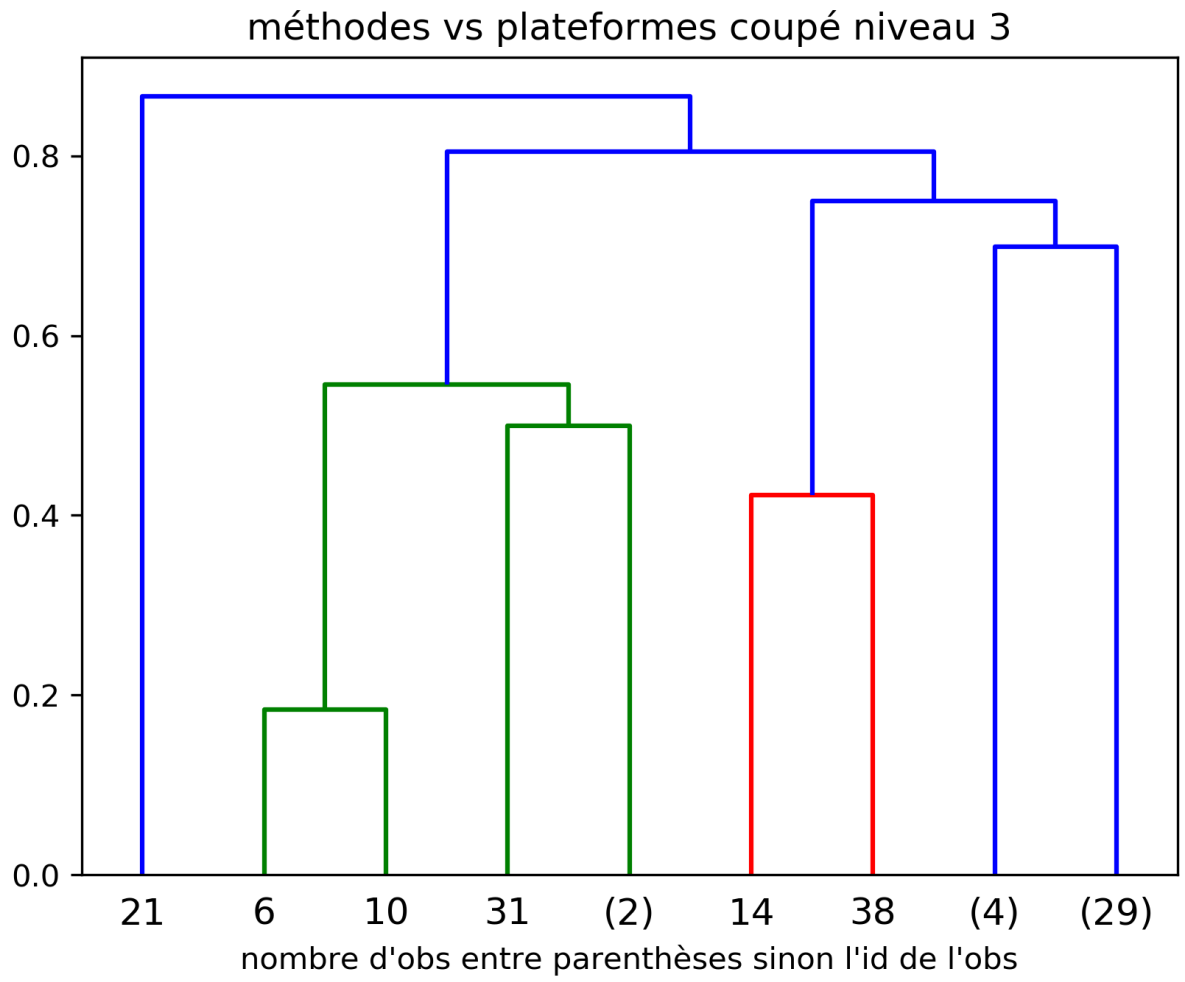


Figure 32 : dendrogramme du clustering méthodes ML et plateformes de calculs, coupé au niveau 3

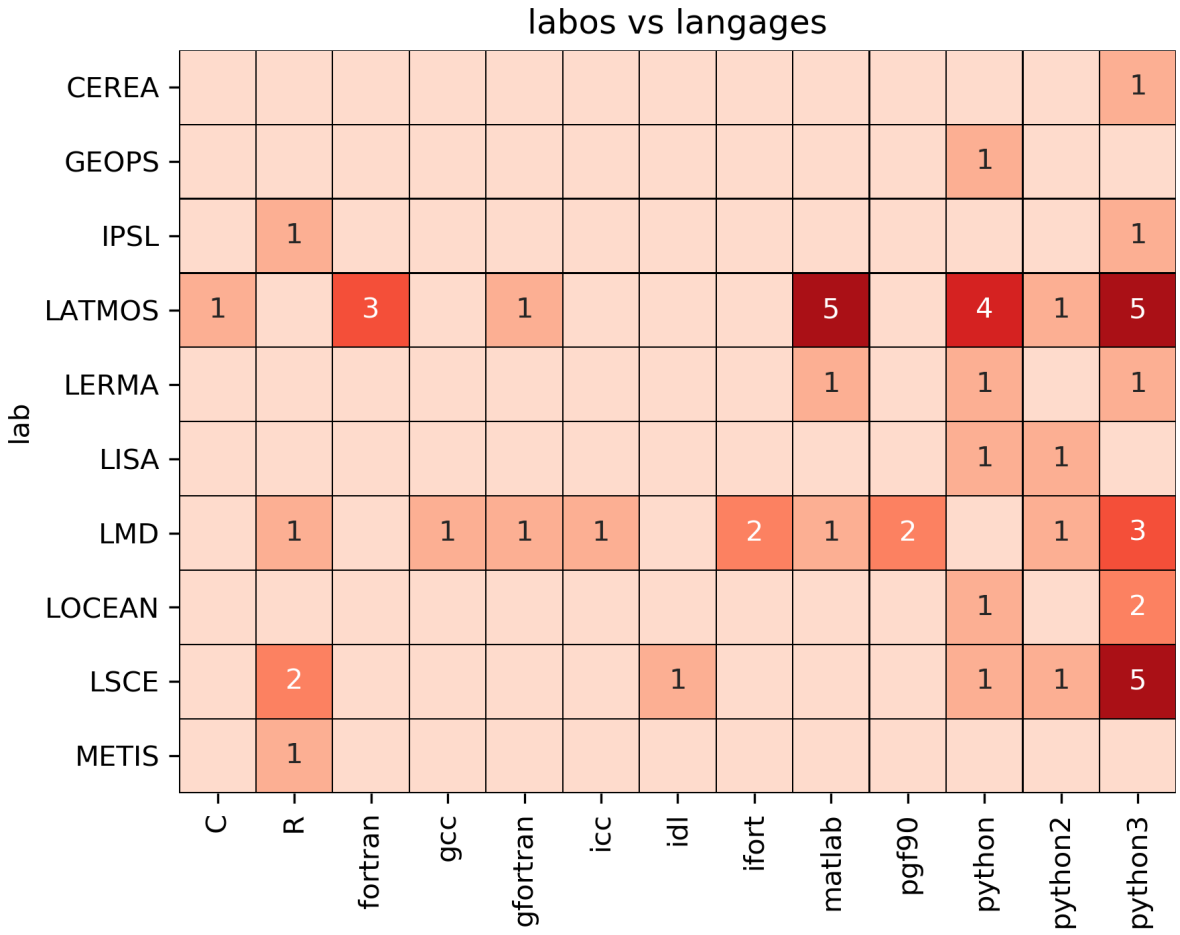


Figure 33 : langages d'implémentation des projets ML par laboratoire

Références

Codd, E. F. (juin 1970). “A Relational Model of Data for Large Shared Data Banks”. In : *Commun. ACM* 13.6, p. 377-387. issn : 0001-0782. doi : 10.1145/362384.362685. url : https://doi.org/10.1145/362384.362685.

Dedić, Nedim et Clare Stanier (2017). “Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery”. In : *Innovations in Enterprise Information Systems Management and Engineering*. Sous la dir. de Felix Piazolo et al. Cham : Springer International Publishing, p. 114-122. isbn : 978-3-319-58801-8.

Fox, Charles (2018). *Data Science for Transport*. Springer. url : https://www.springer.com/us/book/9783319729527.

Laney, Douglas (fév. 2001). *3D Data Management : Controlling Data Volume, Velocity, and Variety*. Rapp. tech. META Group. url : http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Magoulas, Roger et Ben Lorica (2009). *Release 2.0*. Rapp. tech. O’Reilly. url : https://www.oreilly.com/data/free/release-2-issue-11.csp.

Table des figures