PROJET FIL ROUGE - Modélisation Informatique des choix en restauration collective rapport de mi-projet

Jules MARCAIS
Hélène PHILIPPE
Hipolyte DREYFUS
Judith COUTROT
Luis RACCA

Janvier 2023

1 Cahier des charges

Contexte

Les choix alimentaires sont multifactoriels, et il est prouvé que des influences sociales jouent un rôle dans ces choix. En particulier, les choix en restauration collective (qui représentent quasiment une majorité des repas pris hors-domicile), sont très fortement influencés socialement. C'est un système complexe : les choix des autres m'influencent, et inversement, mes choix influencent ceux des autres.
Dans un objectif de santé publique et de science sociale, il est intéressant de connaître les dynamiques sous-jacentes à ces choix, pour les influencer vers des aliments plus sains.

Problématique(s)

On s'intéresse au rôle des influences interpersonnelles (ie comment ce que mange autrui autour de moi influence mes propres choix, et inversement) dans les choix alimentaires que font les individus.
N. Darcel fait l'hypothèse que ces influences interpersonnelles comptent pour beaucoup dans les choix alimentaires. Notre objectif est donc de valider ou réfuter cette hypothèse au cours de ce projet. Nous chercherons à préciser quelles sont ces influences sociales (le fait de manger seul, à plusieurs, connaître le choix des autres...) et leurs poids.

Expérience et données

Données réelles :

1. Résultats de formulaires remplis anonymement par environ 500 personnes sur 6 jours de collecte. Les questions (QCM) dans ce questionnaire portaient sur :

(a) genre
(b) âge
(c) taille
(d) poids
(e) activité physique
(f) régime particulier
(g) végé
(h) fréquence de venue au CROUS
(i) niveau de faim
(j) niveau de stress
(k) connaissance entourage
(l) importance entourage
(m) importance équilibre alimentaire
(n) importance de la quantité
(o) importance aspect
(p) importance de l'attente
(q) convives

2. Listing des choix alimentaires (entrée, plat, dessert, heure) associées grâce au numéro de formulaire pour chacune des personnes ayant rempli le questionnaire. Listing rempli par une personne à la caisse.

Données de simulation:

Grâce à la plateforme GAMA, modélisation des restaurants et de personnes avec des règles de décisions. Nous disposons des inputs du modèle (les paramètres gérant les règles de décision des agents et de l'environnement) et des outputs de la simulation (les choix d' entrées / plat / dessert de chaque agent de la simulation). Les outputs des simulations sont comparés avec les résultats des formulaires. Les paramètres données en input sont modifiées pour faire correspondre au mieux les résultats de la simulation avec les résultats réels.

Besoins en Prétraitement :

L'expérience a été réalisée dans deux restaurants CROUS du plateau de Saclay (le lieu de vie à côté de l'IUT d'Orsay et la Kantine de l'ENS) et à deux périodes différentes (Octobre 202. et Juin 202.) : les données récoltées sont sous différentes formes. Les questions posées lors du questionnaire de sorties diffèrent légèrement. Il faudra donc faire du prétraitement de données pour les faire correspondre et générer un modèle à partir du plus grand nombre d'exemples possibles. Par exemple :

1. Faire correspondre les variables des données de Juin aux variables des données d'Octobre, notamment les horaires de passage à la caisse, les plats choisis, ...

2. Transformation éventuelle des données catégorielles (genre, plats, régime...) en données numériques. Ou éventuellement inversement, transformer des données numériques en catégories (transformer horaire en plages horaires, catégories d'IMC...)

3. Gestion des inputs du modèle multi-agents

Objectifs et calendrier associé

1. Pour le 14 Décembre : Bibliographie, identifier des ressources de l'état de l'art en lien avec le sujet

2. Pour le 1er Janvier : S'approprier les données : Dans un premier temps, on ne considère pas le modèle multi-agents. Nous réaliserons des statistiques descriptives sur les données obtenues dans les cantines

   (a) Commencer par séparer les variables sociales (influence des amis et des autres), personnelles (régime alimentaire, faim, goût moyen en France) et environnementales (perception des plats, temps d'attente).

   (b) Faire des analyses statistiques en fonction des différentes variables

      i. Corrélations/ACP entre les plats (quels plats sont pris ensemble) et entre les variables

      ii. Régression linéaire (un plat en fonction des variables pour voir quelles variables influencent le choix de chaque plat)

   (c) Visualisation du GridSearch : affichage des valeurs de fitting des modèles en fonction des valeurs associées aux variables (heatmap avec 2 ou 3 variables à la fois)

3. Pour le 18 Janvier : Modifier le modèle pour voir quels descripteurs sont les plus importants : N. Darcel compte sur le fait que nous trouvions des descripteurs plus pertinents.

4. Pour le 20 Février : Évaluer l'importance des influences interpersonnelles au sein de ce modèle

2 Préprocessing

Avant de pouvoir travailler sur les données il a fallu les standardiser. En effet, la retranscription numérique des formulaires présentaient des erreurs de frappe qu'il fallait corriger. De plus, pour pouvoir travailler plus facilement, nous avons choisi d'encoder les données sous format "one-hot".

A. data cleaning

En digitalisant les formulaires, les mots n'ont pas étés écrits exactement de la même manière, de telle sorte que certains plats sont considérés comme différents alors que ce n'est pas le cas. La figure 1 donne un aperçu de ce phénomène. On peut observer que le plat "Boeuf Bourguignon + Torsades + Poêlée brocolis" apparaît 3 fois dans la liste des plats uniques, comme si c'était 3 plats différents. Ceci est dû à la présence d'espaces non uniformes.

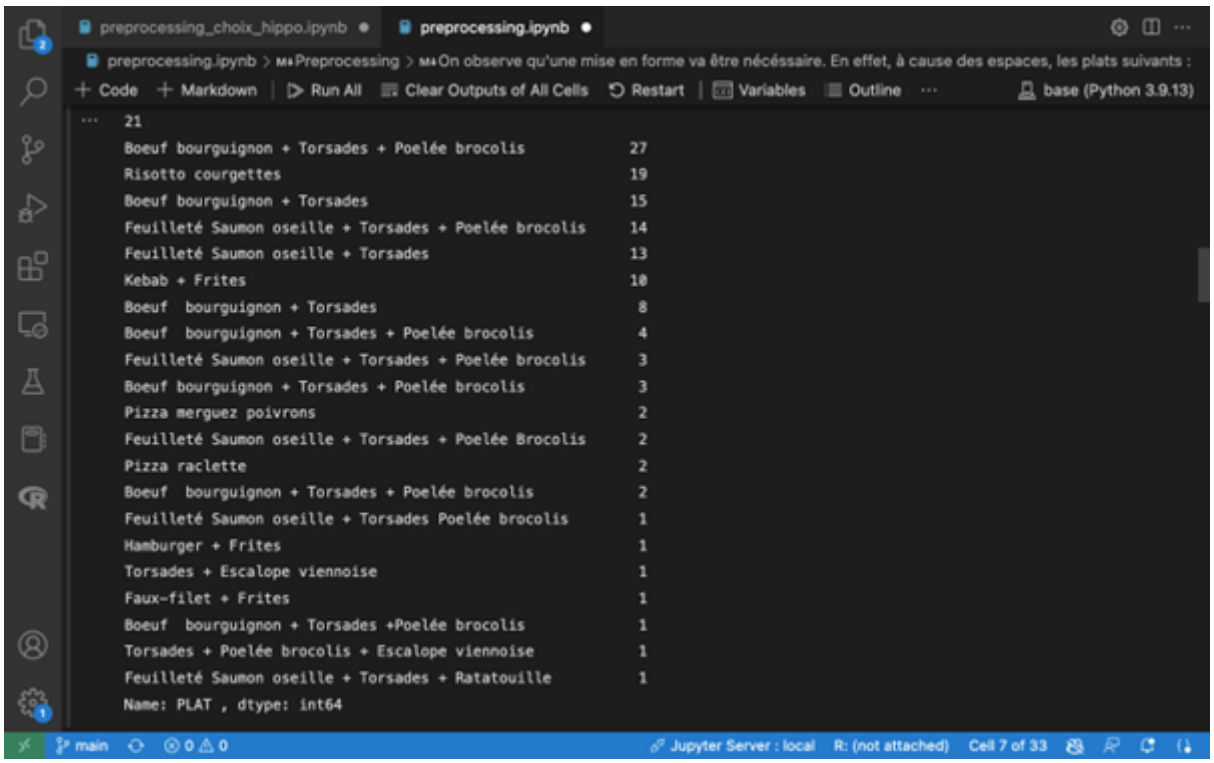

Figure 1 : Illustration des erreurs de frappe lors de la transcription des résultats des formulaires.

Pour régler ce problème, nous avons supprimé tous les espaces. De même, pour résoudre d'autres genres de fautes de frappe (dues à la présence/absence d'accents ou de majuscules, par exemple) nous avons supprimé les accents et placé toutes les lettres en minuscules. Enfin, certaines coquilles ne pouvaient pas se résoudre par une standardisation automatique avec python, mais nécessitaient que cela soit fait manuellement directement sur les fichiers csv. Par exemple, sur la figure 1, on observe qu'un "+" a été oublié entre torsades et brocolis sur la quinzième ligne : "Torsades Poêlée brocolis" est alors considéré comme un nouveau plat. Il faut donc rajouter manuellement le "+" dans le dossier csv. Pour ne pas corrompre les fichiers originaux que nous avaient transmis M. Darcel, nous avons choisi d'effectuer ces modifications sur une copie des fichiers. Nous avons procédé ainsi sur l'ensemble des fichiers csv récapitulant les choix de plats. Le jupyter notebook associé est nommé "cleaning.ipynb". Les fichiers csv contenant les données nettoyées sont dans le dossier data_clean.

B. Data preprocessing

À partir de ces données nettoyées, nous avons choisi de modifier l'organisation du fichier csv. Nous avons transformé les colonnes PLAT, ENTRÉE et DESSERT en trois colonnes chacune, de telle sorte que la composition des plats soit prise en compte, plutôt que de considérer le plat comme une entité monolithique. Ceci présente l'intérêt suivant : les personnes ayant choisi "boeuf + torsades" présentent un profil semblable à une personne ayant choisi "boeuf + torsades + brocolis". Or si on ne considère qu'une colonne (donc un plat monolithique), ces deux compositions d'assiette sont considérées comme deux plats tout aussi différents que ne le sont "boeuf bourguignon + torsades" et "saumon à l'oseille + brocolis". Ceci est réalisé grâce au jupyter notebook "split_columns.ipynb".

Enfin, nous avons binarisé les variables : on crée autant de colonnes que de proposition d'aliments, et on indique 1 si la personne l'a choisi, 0 sinon. Ceci est réalisé avec le jupyter notebook "one_hot.ipynb" et les résultats de cet encodage sont enregistrés dans le dossier du même nom. C'est sur ces données que nous avons ensuite travaillé.

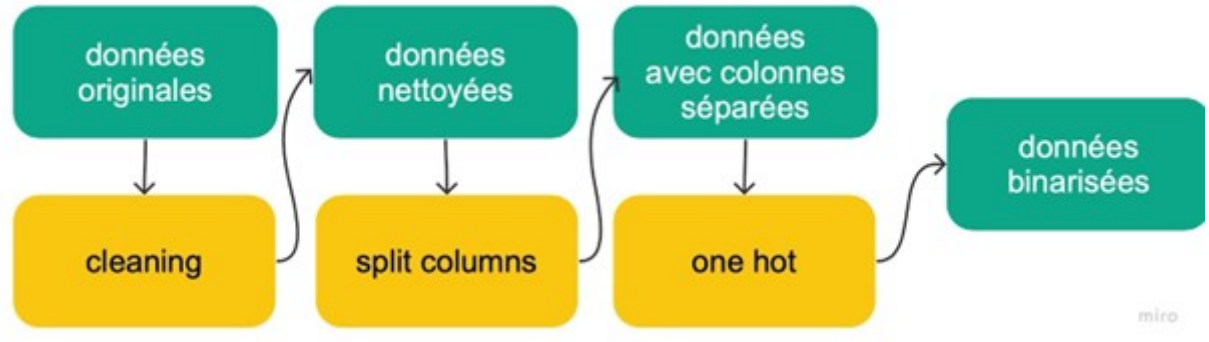

Figure 2 : Pipeline du preprocessing des données

3 Satistiques descriptives

Pour avoir un regard global sur les données des formulaires et des choix, nous avons commencé par faire des statistiques descriptives. Pour ce faire, nous avons utilisé R dans le fichier "Stat_R.Rmd". Nous avons traité les variables des données formulaires, qui seront par la suite parfois aussi appelées variables explicatives, telles que l'activité physique, l'âge, la taille, la faim, l'influence sociale perçue etc. des individus interrogés. Ces variables sont explicatives car ce sont celles qui, dans le modèle, vont permettre de d'expliquer les choix alimentaires des individus. L'autre type de données traitées et décrites ici seront donc les variables à expliquer : les choix de plats des individus interrogés.

A. Analyse Univariée

Dans un premier temps, nous avons fait une analyse univariée des variables pour avoir un visuel sur les données récoltées indépendamment les unes des autres et mettre éventuellement en évidence des disparités dans leurs répartitions.

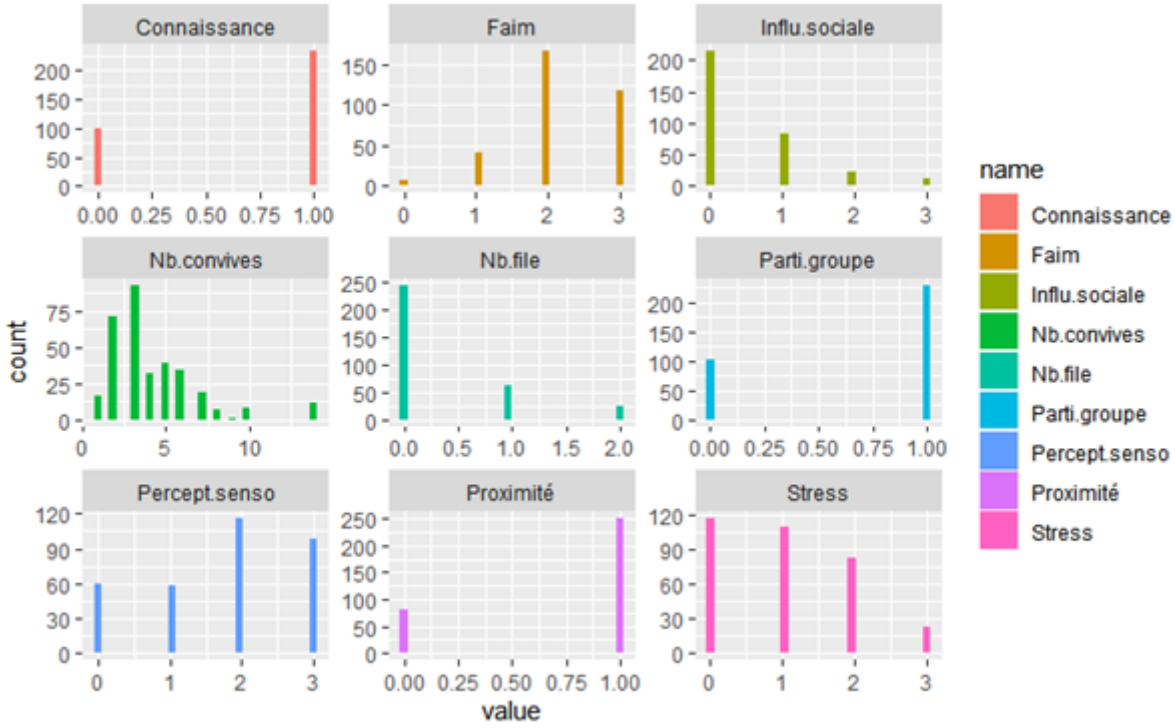

Figure 3 et 4 : Histogrammes des variables explicatives numériques des données issues des formulaires

Les variables sont codées de trois façons différentes, issues du preprocessing :

1. Les variables âge, IMC, nombre groupes entrés, nombre convives, poids et taille initialement numériques sont laissées telles quelles et on peut en observer la répartition ci dessus.

2. Les variables récoltées sous forme d'échelle sont normalisées (centrée, réduite). Certaines sont assez réparties comme le stress, l'activité physique ou la perception sensorielle. D'autres comme la faim ne sont pas réparties le long des abscisses : c'est le cas de la faim où les faibles valeurs sont rares.

3. Les variables qui dans les formulaires étaient sous forme OUI/NON ont été codées en 0/1, pour avoir des variables numériques.

Souvent une des valeurs est bien plus représentée que l'autre. C'est le cas pour le régime végétarien et les autres régimes particuliers. La variable Genre est catégorielle pour le moment. Nous envisageons de la transformer en numérique si le besoin se présente dans la suite de notre étude.

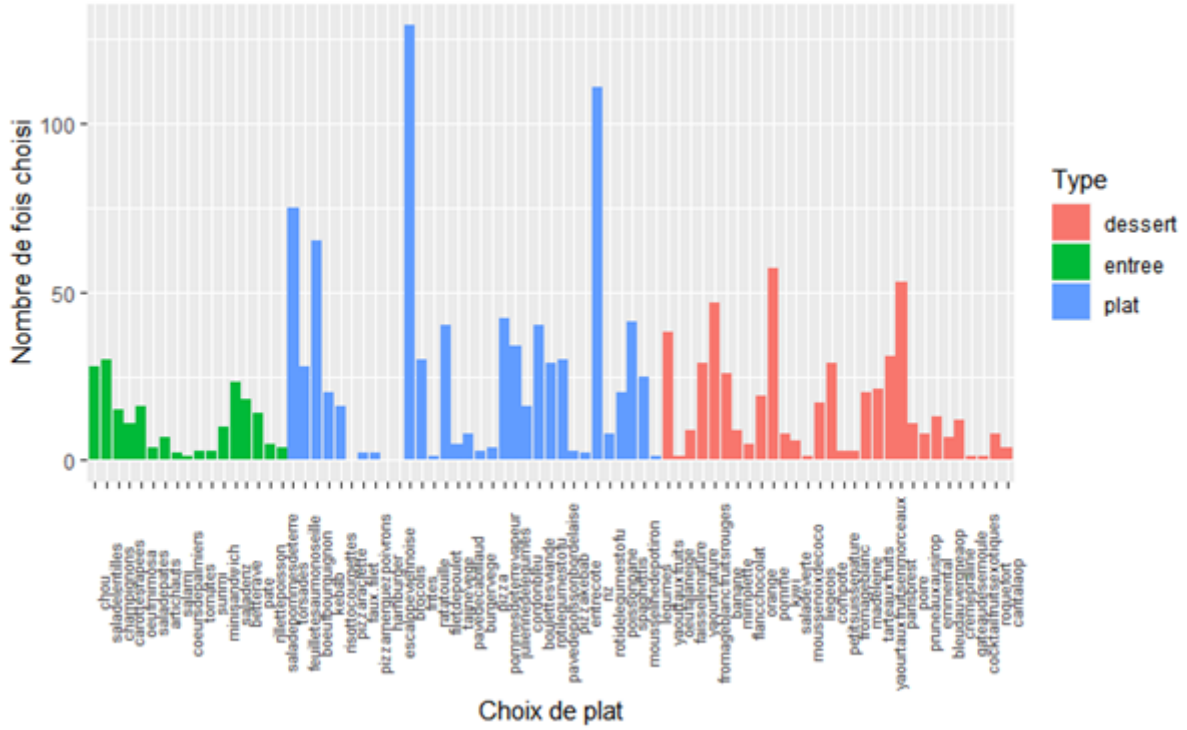

Figure 5 : Barplot des choix des individus à expliquer

Pour représenter les données de choix des plats, le graphique ci-dessus indique le nombre de fois où certains plats, entrées ou desserts ont été choisis, au cours d'une des études. Il permet de voir que certains plats sont beaucoup plus choisis que d'autres, il faudra prendre cette disparité en compte lors de l'apprentissage d'un modèle pour éventuellement pondérer ou regrouper les choix alimentaires tels qu'ils l'ont été dans le fichier "Regroupements_plats-version-article2022.csv".

B. Analyse en Composantes Principales

Dans un second temps, nous avons étudié les variables en interactions les unes avec les autres. Pour ce faire nous avons calculé avec R la matrice de corrélation de entre variables explicatives du formulaire d'une part et variables de choix à expliquer d'autre part.

B.1 - Étude des corrélations entre les variables quantitatives du jeu de données.

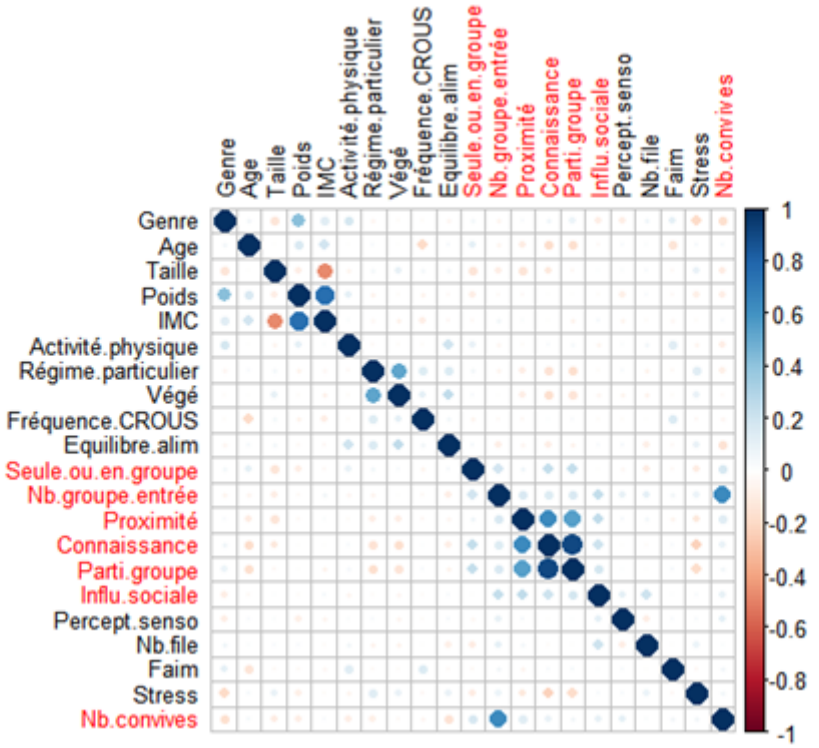

Figure 6 : Matrice de corrélation des variables explicatives des formulaires

D'une manière générale, les variables quantitatives sont peu corrélées. Il est toutefois possible de noter des corrélations et anti-corrélations auxquelles il était possible de s'attendre: Taille, IMC et Poids, ou encore Végétarien et Régime particulier. Il sera peut être intéressant par la suite de s'affranchir de certaines de ces variables qui n'apportent pas plus d'informations les unes par rapport aux autres. Par ailleurs, on observe des corrélations entre les variables liées à l'influence sociale (marquées en rouges), qui sont liées dans cette représentation par des points bleus, plus ou moins marqués.

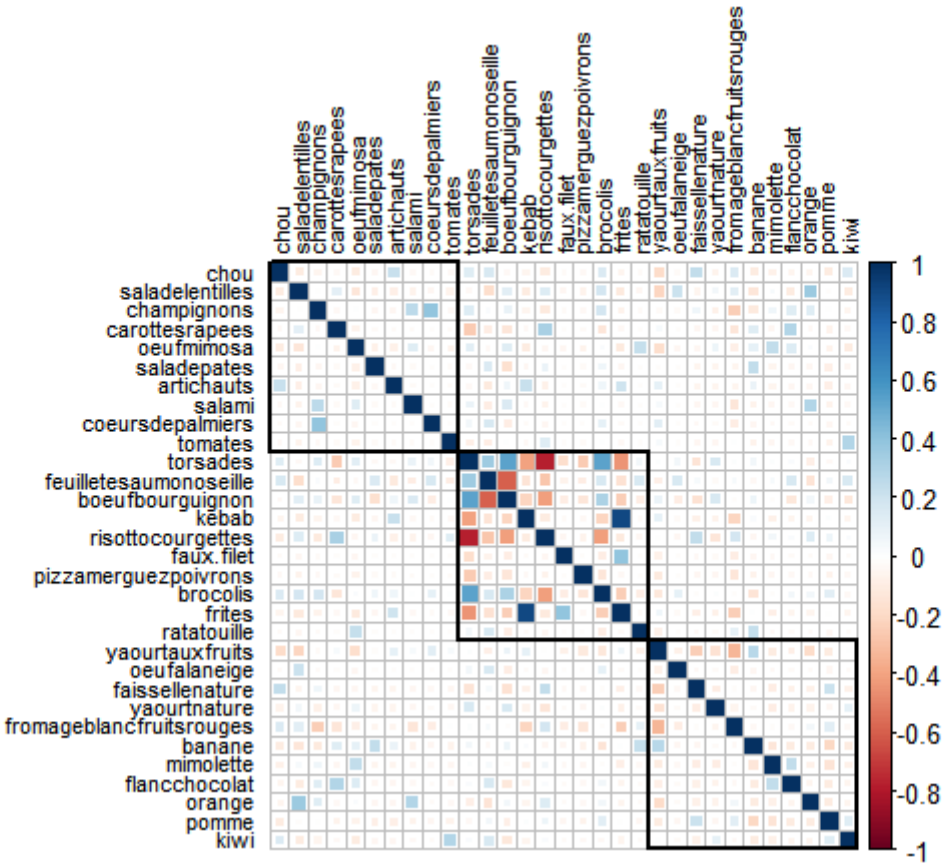

Figure 7 : Matrice de corrélation des variables de choix à expliquer

Pour plus de lisibilité, cette matrice de corrélation ne porte que sur les aliments possibles d'un seul des repas étudiés (21 Octobre).

Les rectangles représentés en noir montrent respectivement (haut-gauche, centre, bas-droite) les corrélations au sein des types 'ENTRÉES', 'PLAT' et 'DESSERTS'.

On observe de fortes corrélations et anti-corrélations au niveau des plats, qui correspond au fait qu'en restauration collective, pour des raisons de praticité et de menus, certains éléments du plat principal ne sont servis qu'avec certains accompagnements.

De faibles corrélations et anti-corrélations sont par ailleurs indiquées dans la représentation de matrice ci-dessus. Il en existe aussi inter-type : entre une entrée et un dessert par exemple. Les variables explicatives telles que la faim où l'équilibre alimentaire pourraient expliquer ces corrélations. C'est pourquoi il sera intéressant d'avoir ces variables explicatives qui ne sont pas directement liées à l'influence sociale, à prendre en compte dans le modèle.

B.2 - Réalisation de l'ACP

On s'intéresse à la contribution de chaque variable explicative au choix de l'individu (variable à expliquer).

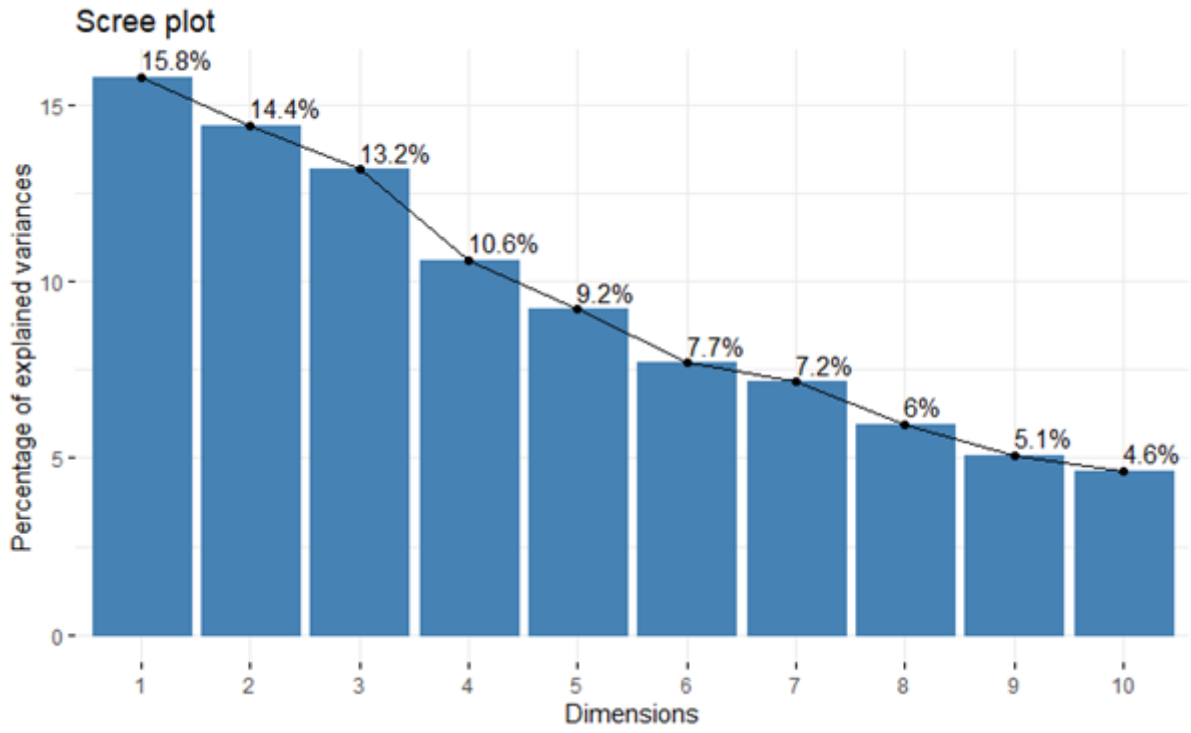

Figure 8 : Pourcentage de variance expliqué par chaque dimension

Nous n'observons pas deux ou trois dimensions expliquant une grande majorité de la variance (en effet, pour expliquer 90% de la variance, il faut utiliser 8 dimensions). Avec 3 dimensions, seuls 43% de la variance sont expliqués, ce qui n'est pas suffisant. Cette méthode ne nous permet pas de décider du nombre d'axes principaux qui suffisent pour trouver des profils intéressants dans les données.

Pour autant, nous avons projeté le nuage des variables sur les deux premières dimensions, puis sur la première et la troisième dimension. Ces cercles sont plutôt cohérents puisque les variables "nombre de convives", et "nombre groupe entré" sont proches et plutôt bien représentées (cos^2 élevé), ainsi que les variables Poids et IMC pour la projection sur les dimensions 1 et 2. De même, la projection sur les dimensions 1 et 3 (et sur les dimensions 3 et 5) montre des variables proximité et influence sociale relativement proches mais pas assez bien représentées dans ces dimensions pour pouvoir conclure (cos^2 peu élevé).

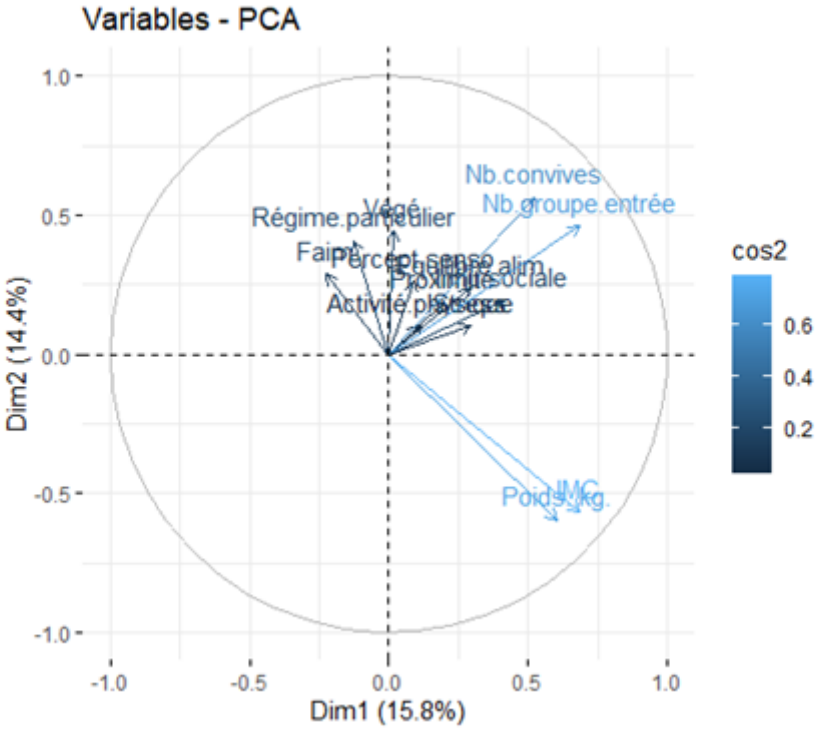

Figure 9 : Cercle des corrélations et projections des variables sur les dimensions 1 et 2

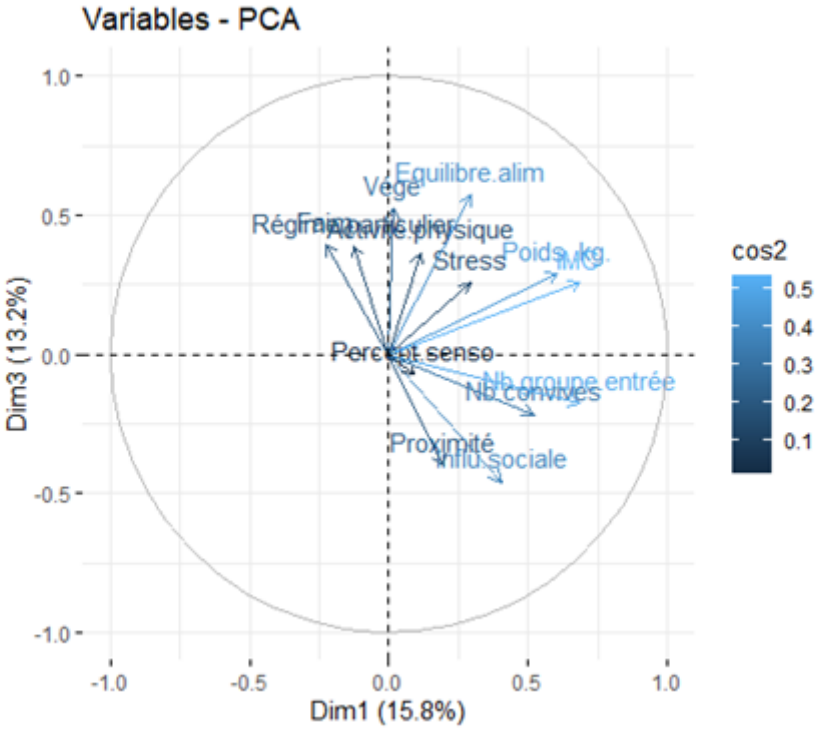

Figure 10 : Cercle des corrélations et projections des variables sur les dimensions 1 et 3

C. Conclusion sur les statistiques descriptives

Pour l'instant, les variables traitées directement liées à l'influence sociale sont la perception de l'influence sociale par l'individu, le nombre de convives attablés, le nombre de personnes dans le groupe entré dans le self en même temps, la connaissance et la proximité au sein du groupe attablé.

Certaines variables n'ont pas encore été prises en compte, notamment les

variables “Ami” qui correspondent aux numéros de formulaires des convives ayant déjeuné ensemble. Par la suite, nous souhaitons étudier statistiquement ces variables en créant des groupes de personnes, chacun des ces groupes étudié séparément.

4 Modèle multi-agents

Les systèmes multi-agents sont des architectures informatiques qui utilisent des agents autonomes pour modéliser des interactions complexes entre différentes entités. Ils simulent la dynamique des interactions entre des agents rationnels, prenant des décisions et interagissant les uns avec les autres. Les systèmes multi-agents sont particulièrement pertinents à utiliser en sciences sociales pour étudier le choix d'individus dans des groupes car ils permettent de mieux comprendre la façon dont un individu, ou un groupe, se comporte face à une série de facteurs et de circonstances. En modélisant les agents avec des caractéristiques personnelles telles que l'âge, le sexe, le poids... Cela permet de modéliser ce à quoi ressemble la vie réelle et les relations entre les individus dans un groupe et un environnement donné. De plus, les systèmes multi-agents sont capables de simuler l'interaction de milliers de personnes, ce qui permet une analyse beaucoup plus précise des données et un meilleur aperçu des choix faits par chaque individus en fonction de ses caractéristiques pour le mettre en regard du comportement global du groupe.

Dans le cadre de ce projet, le système multi-agent mis en place par N. Darcel et son équipe vise à simuler une foule dans une réplique d'un restaurant universitaire pour étudier et tenter de faire apparaître l'influence sociale entre les individus dans le cadre des choix alimentaires en restauration collective. Chaque agent est modélisé avec un certain nombre de variables dans lesquelles on peut retrouver l'âge, l'IMC, la propension à être influencé par les autres, son appétences pour certains plats... et de caractéristiques et fonctions, comme une zone de détection autour de lui pour “percevoir” ce que les autres agents dans son environnement proche consomment ou encore des fonctions de décision pour sélectionner le prochain plats à aller chercher et le chemin le plus court pour y parvenir.

De premiers résultats encourageants ont pu être obtenus grâce à un premier MAS. On peut voir ci-dessous la manière dont ces résultats peuvent se présenter, en vert les données réelles et en rouge les données de simulation (Figure 11) :

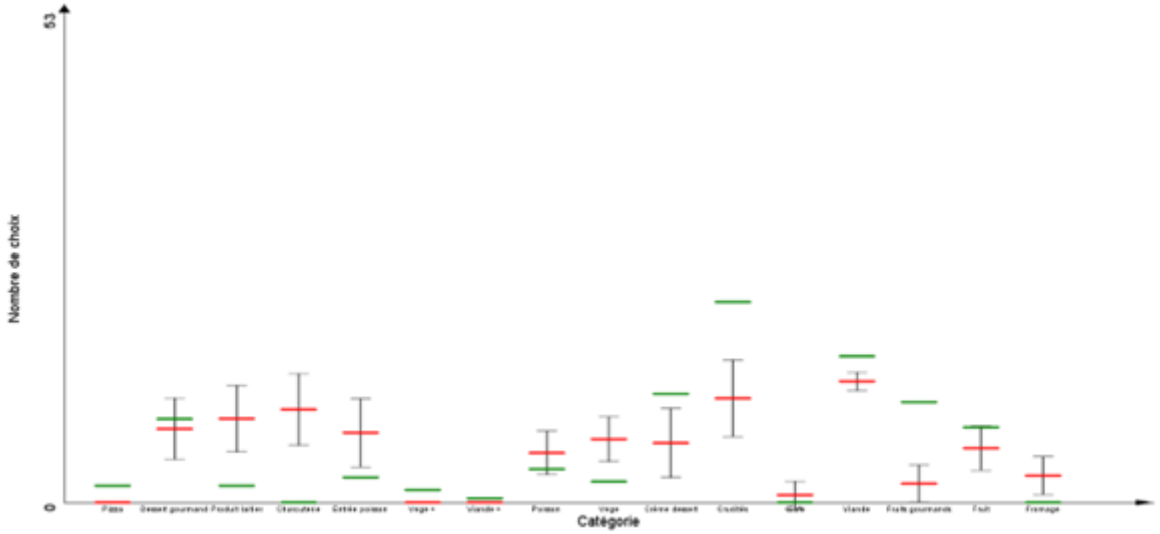

Figure 11 : Comparaison des performances du modèle à prédire le nombre de plats consommés contre le nombre de plats réellement consommés

Pour la suite de l'étude, nous souhaitons comparer les résultats des simulations réalisées avec le modèle multi agent avec les résultats observés. Pour l'instant nous avons seulement eu le temps de nous familiariser avec la plateforme GAMA. C'est un logiciel open source permettant de réaliser des simulations de systèmes multi-agents spatiaux-dynamiques. Chaque simulation se présente comme suit, avec un panneau dynamique (Figure 12) permettant la visualisation de la motricité des agents et un autre nous permettant de suivre en temps réel l'évolution des choix de la population (Figure 11) :

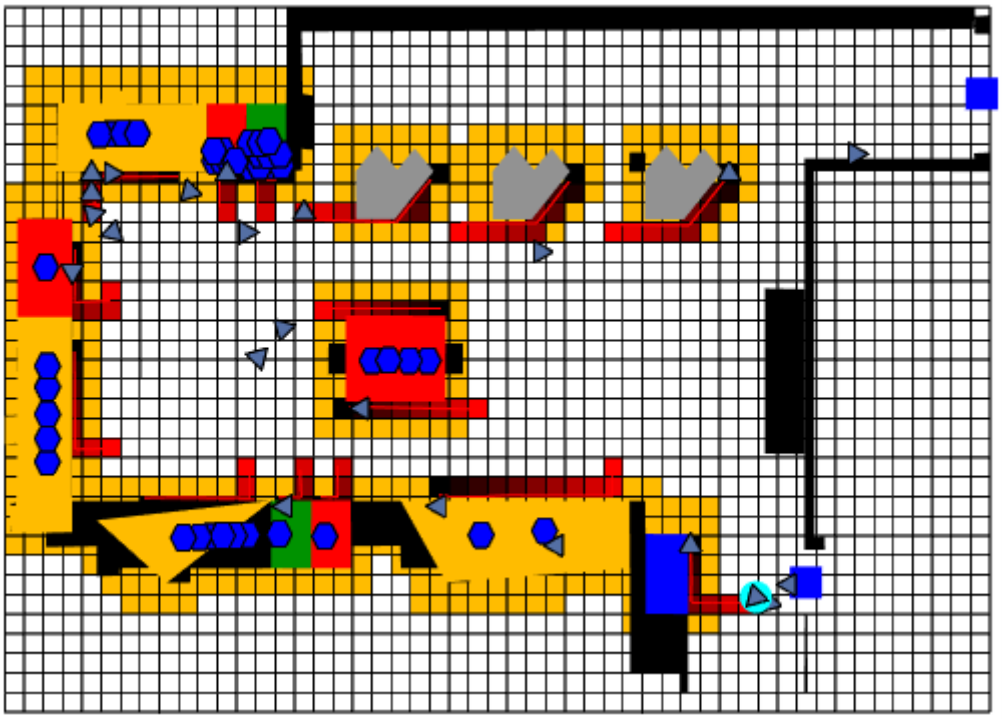

Figure 12 : Représentation spatiale du restaurant universitaire. Hexagones bleus = plats, Triangle bleus = Agent, Formes Grises = Caisses

5 Discussion / pistes d'améliorations

Nous disposons de plusieurs leviers d'amélioration pour la suite de ce projet.

Le premier consiste à modéliser statistiquement les résultats. En effet, bien que le modèle permette de prédire de manière relativement correcte les quantités de chaque plat consommées par rapport aux données réelles, nous aimerions pour la suite essayer de proposer d'autres approches moins techniques que le modèle multi-agents. Pour confirmer ou infirmer l'effet de l'influence sociale, nous allons par exemple étudier si les plats consommés par les amis d'un individu permettent de prédire mieux que le hasard ce que cet individu va consommer. Pour ce faire, nous utiliserons simplement des tirages aléatoires qui définiront une composition de plateau. On la compare avec les plateaux réellement observés. Puis, on réitère l'experience, mais cette fois ci avec une pondération des tirages aléatoires selon les plateaux des convives. On espère observer une meilleure prédiction de plateaux à partir de ceux des convives que celui obtenu avec un tirage parfaitement aléatoire.

Le second levier dont nous disposons est l'amélioration du système multi-agents pour rendre le comportement des agents et les paramètres d'environnement plus proches de la réalité. Par exemple, les agents présentent une zone de sensibilité qui leur permet d'être influencés par ce qui se trouve près d'eux. Cette zone est modélisée dans le modèle actuel comme un cercle autour de chaque agent, alors qu'il serait plus pertinent de modéliser cette zone comme un cône pour imiter la vision.

De plus, le modèle ne prend pour l'instant pas en compte l'épuisement des stocks de produit dans la réalité. Là où un restaurant peut subir une rupture de stock, ce n'est pas le cas dans la simulation, ce qui tend donc à favoriser certains plats plus appétents que d'autres au fil du temps. Pour améliorer le modèle, nous pouvons implémenter un stock pour chaque plat à partir de ceux que les restaurants universitaires proposent en général. Si nous ne parvenons pas à accéder à ces données, l'autre possibilité consiste à utiliser l'horodatage. On peut implémenter dans la simulation une rupture de stock après un certain nombre d'itérations à partir des données rélles. Pour inférer la rupture d'un produit dans les données rélles, on considérera que s'il n'apparait plus dans les plateaux après une certaine heure, il est en rupture.

Enfin, des statégies de sélection de variables ont été envisagées pour tenter de faire ressortir celles qui influent le plus les choix des individus, de manière à quantifier le poids de chaque variable. Cependant, il nous reste encore à identifier la ou les méthodes les plus pertinentes qui nous permettraient de réaliser ceci sur les données dont nous disposons.