Calibration de simulations multi-agents à l'aide d'une méthode semi-automatique d'analyse du comportement

Kévin Darty, Julien Saunier, Nicolas Sabouret

Calibration de simulations multi-agents à l'aide d'une méthode semi-automatique d'analyse du comportement

K. Darty[a,b] kevin.darty@limsi.fr
J. Saunier[c] julien.saunier@insa-rouen.fr
N. Sabouret[b] nicolas.sabouret@limsi.fr

[a]IFSTTAR, France

[b]LIMSI-CNRS, UPR 3251, Univ. Paris-Sud, Orsay, France

[c]LITIS, INSA de Rouen, France

Résumé

L'un des principaux problèmes en simulation multi-agent est la définition des paramètres du modèle agent et leur calibration. Ce problème est encore plus difficile lorsqu'on considère des environnements virtuels immersifs, dans lesquels les agents intelligents doivent reproduire des comportements humains et apparaître « réalistes » aux yeux des utilisateurs. Dans cet article, nous proposons d'enregistrer et d'analyser les comportements des agents pour évaluer leur similarité avec ceux des humains dans un environnement virtuel immersif. Nous utilisons des méthodes de classification pour construire une abstraction des comportements individuels. Les types de comportement sont étudiés selon la composition de ces classes pour déterminer les manques, capacités et erreurs dans le modèle agent. Cette méthode nous permet 1) d'écarter les jeux de paramètres invalides, 2) de calibrer des simulations valides et 3) d'expliquer les manques du modèle agent pour l'améliorer.

Mots-clés : *Simulation multi-agent, Calibration*

Abstract

In the context of agent-based simulation, a major issue is to define relevant parameters of the agent model and calibrate them. This issue is yet harder in immersive virtual environments, where intelligent agents reproduce human behaviour and interact with users. In this paper, we propose to log and analyse agents behaviour to evaluate their similarity to humans behaviour in an immersive virtual environment. Clustering is then used to get an abstraction of individual behaviours. The behaviour archetypes are studied in terms of cluster members in order to identify agent lacks, capacities and errors. This study enables to 1) dismiss invalid parameter sets, 2) calibrate valid simulations and 3) explain lacks in the agent models for further improvement.

Keywords: *Multi-agent simulation, Calibration*

1 Introduction

Dans le contexte des simulations à base d'agents (voir par exemple [14, 1, 7]), les agents doivent souvent reproduire des comportements similaires à ceux qu'adopteraient des humains dans la même situation. Une difficulté de ces modèles de simulation est d'identifier les jeux de paramètres qui permettent d'obtenir des comportements « valides ». D'une part, le comportement des agents doit être crédible, *i.e.* les valeurs des paramètres doivent conduire à des comportements qu'un humain pourrait adopter. De l'autre, l'ensemble des comportements produits par les agents doit être représentatif de la population simulée, *i.e.* les jeux de paramètres doivent permettre d'obtenir une certaine variabilité dans les comportements des agents.

La majorité des travaux de recherche dans le domaine s'intéressent surtout au niveau individuel. Cela conduit à définir des jeux de paramètres qui correspondent à des comportements moyens ou normatifs. Si ce choix est pertinent pour des simulations de niveau macroscopique, les *SMA* s'intéressent généralement à des phénomènes microscopiques pour lesquels les comportements normatifs ne sont pas adaptés [12, 1]. Dans ce contexte, plusieurs méthodes ont été proposées pour l'analyse semi-automatique des comportements d'agents en fonction des paramètres du modèle. Par exemple, Taillandier *et al.* [17] proposent une méthode pour évaluer les ensembles caractéristiques d'une méthode d'apprentissage supervisé qui contrôle un *SMA* géographique. Caillou *et al.* [2] proposent un modèle de classification des agents suivant leur

comportement pour étudier l'impact des paramètres sur la dynamique du *SMA*. Plus récemment, nous avons proposé dans [5, 4] une méthode semi-automatique d'analyse des comportements des agents dans une simulation immersive en comparant des classifications de comportements produits par des agents et par des humains mis dans la même situation. La crédibilité du comportement des agents est alors étudiée en termes de capacités à reproduire des comportements humains, de manques (*i.e.* de comportements non reproduits) et d'erreurs (*i.e.* de comportements produits par des agents mais aucun humain). Toutefois, aucun de ces modèles ne propose de méthode pour calibrer les paramètres de chaque agent de la simulation à partir de l'analyse des comportements.

Dans cet article, nous proposons d'étendre notre approche proposée dans [5] pour l'évaluation et la calibration de simulations multi-agents. Pour cela, nous analysons les comportements produits par les agents indépendamment du modèle sous-jacent. Nous considérons donc les agents comme des boîtes noires et collectons des données sur leur comportement en fonction des paramètres du modèle. Nous comparons alors les traces de simulation à l'aide de mesures calibrées sur les comportements produits par les humains en simulation participative. L'agrégation des traces de comportement à l'aide de méthodes de classification permet d'obtenir des archétypes de comportement [4]. La composition de chaque classe (agents, humains ou mixte) permet d'identifier et d'expliciter les comportements agents en fonction des paramètres. Enfin, les classes nous permettent de définir la population d'agents pour la simulation.

La prochaine section présente les travaux connexes dans le domaine de l'évaluation des comportements, qui est à la base de notre approche, et nous présentons brièvement la méthode de classification sur laquelle nous nous appuyons. La section 3 présente notre méthode d'évaluation des comportements et de calibration des paramètres. La section 4 montre la mise en œuvre de cette méthode dans le contexte de la simulation de conduite.

2 Travaux connexes

2.1 Analyse du comportement

La notion de comportement couvre différents aspects, des actions bas niveau sélectionnées par un agent lors d'un cycle d'exécution, à des éléments plus complexes comme les mouvements de foule [1] ou les décisions macro-économiques [14]. Cependant, tous ces domaines partagent un même objectif : produire des résultats valides à l'aide de simulations multi-agents pour l'analyse et la prédiction du comportement humain.

La validité est vue dans la majorité des modèles à un niveau macroscopique : les méthodes d'analyse statistique de données sont alors bien adaptées pour déterminer la validité de la simulation [8, 1]. Il s'agit de vérifier, à travers des données quantitatives, que les agents se comportent de manière similaire à ce qui est observé dans une situation « réelle ». Cependant, le fait que le comportement global soit valide ne garantit pas que les comportements individuels soient réalistes. C'est pourquoi, dans nos travaux, nous nous intéressons au réalisme du comportement au niveau microscopique : chaque agent devrait adopter un comportement ressemblant à celui d'un humain, tout en maintenant la cohérence au niveau macroscopique.

La comparaison de traces au niveau microscopique se heurte à un problème important [2] : les données recueillies ne peuvent être analysées directement puisque celles-ci sont souvent bruitées et de nature temporelle, alors que les comportements recherchés sont de plus haut niveau. Ceci implique un recours à des traitements de données de façon à donner un sens aux traces de bas niveau.

Dans le domaine des agents virtuels, plusieurs travaux se sont intéressés à la crédibilité des agents et à leur ressemblance aux humains du point de vue de leur réaction affective [3], du comportement non verbal [15] ou de la décision [1]. Ces méthodes s'appuient sur l'évaluation de la crédibilité du comportement des agents par un observateur externe [13]. Bien adaptées pour étudier des agents virtuels, elles ne peuvent pas être utilisée pour traiter des grands nombres d'agents, comme ceux que nous rencontrons dans les simulations multi-agents. En pratique, il est impossible de traiter et tester tous les paramètres sur des centaines d'agents via des méthodes qui nécessitent que le comportement de chaque agent soit observé et analysé par plusieurs humains. Il existe peu de travaux qui s'intéressent à l'analyse objective (par opposition aux approches subjectives à base de jugement humain) et automatique. Des travaux [6, 11] proposent des approches à base d'apprentissage automatique selon des variables de bas niveau, alors que Caillou *et al.* [2] pro-

posent de définir, avec l'aide d'experts du domaine, des variables de haut niveau qui sont ensuite utilisées pour décrire les comportements analysés via un algorithme de classification automatique.

La principale limite de ces approches objectives est que, si elles permettent d'identifier les différences entre des catégories de comportements extraites à partir des traces de simulation (nous parlerons de *classes de traces*), elles ne fournissent aucune information au-delà des variables de bas niveau qui ont été utilisées. En particulier, elles ne permettent pas de donner un sens aux classes obtenues : les comportements détectés restent implicites. Au contraire, les approches subjectives, parce qu'elles s'appuient sur des analyses de haut niveau faites par des humains à travers des questionnaires validés par des experts, permettent d'obtenir des classes de comportement. Le modèle présenté dans [5] propose de combiner les approches objectives et subjectives pour tirer parti des deux méthodes. Nous présentons brièvement son fonctionnement dans la prochaine section.

2.2 Approche mixte

La méthode présentée dans [4] permet d'évaluer les comportements des agents dans le contexte d'environnements virtuels (*EV*) immersifs, en combinant l'analyse des traces d'interactions (approche objective) et les annotations d'observateurs (approche subjective). Les données de la simulation sont classifiées en classes de traces. Les questionnaires permettent de définir des catégories d'utilisateurs pour caractériser les comportements des humains et des agents. Nous évaluons alors le comportement des agents en étudiant la composition des classes de traces et en les comparant aux catégories de comportement des humains.

Cette méthode comprend 5 étapes :

1. la collecte des données de comportement des agents –virtuels– dans la simulation ;
2. la collecte des données de comportement des participants –humains– dans la même situation à l'aide de simulations participatives ;
3. l'annotation de ces données par des observateurs humains (appelés annotateurs) ;
4. le traitement des données et la classification automatique, ce qui conduit à des classes de comportements d'humains et d'agents ;
5. la comparaison des classes, *i.e.* l'analyse de leur composition et l'explicitation du comportement.

La principale limite de cette approche est que le résultat de l'analyse ne permet pas de réévaluer les paramètres de la simulation pour améliorer la validité des comportements produits. Dans cet article, nous proposons d'étendre la dernière étape de cette méthode de telle sorte que le résultat de l'évaluation puisse permettre la calibration du modèle de simulation et l'explication, en termes de valeurs de paramètres, des capacités et des manques dans les comportements d'agents.

3 Analyse & calibration

Nous proposons d'étendre la méthode précédente de la manière suivante (voir la figure 1) :

— nous utilisons l'étape d'abstraction pour calculer des scores qui mesurent d'une part la proportion de capacités, d'erreurs et de manques dans les comportements produits par les agents, et d'autre part la proportion d'agents qui reproduisent des comportements humains, en prenant en compte la fréquence d'occurrence de ces comportements dans les traces de simulation ;

— nous établissons une corrélation entre ces comportements corrects et les paramètres des agents pour calculer une nouvelle distribution de jeux de paramètres ;

— nous appliquons cette méthode en boucle pour tester les nouveaux paramètres et les évaluer en termes de niveau de reproduction de comportements humains. En cas de comportements manquants, nous explorons l'espace des paramètres, et si le modèle d'agent peut être modifié, nous corrigeons les erreurs ou complétons les comportements des agents.

3.1 Abstractions et scores

La dernière étape de la méthode initiale consiste à comparer le comportement des agents avec celui des humains en analysant la composition des classes. Trois types de classes peuvent être identifiés :

1. les classes contenant à la fois des humains et des agents : elles correspondent à des comportements de haut niveau correctement reproduits par les agents. Nous notons $\mathbb{C}_M$ l'ensemble de ces *classes mixtes*.

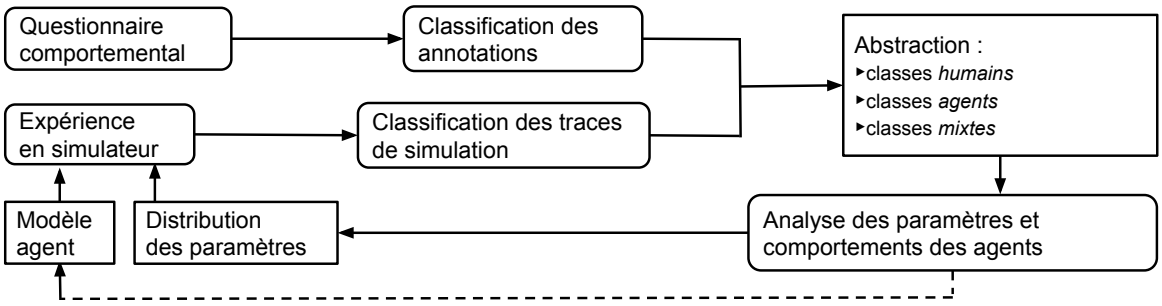

FIGURE 1 – Etapes de la méthode d'analyse et de calibration des comportements.

2. les classes contenant uniquement des agents : elles correspondent à des comportements qui ont été produits seulement par des agents. Dans la majorité des cas, cela correspond à des erreurs dans la simulation, mais cela peut aussi provenir d'un trop petit nombre de participants humains. Nous notons $\mathbb{C}_A$ l'ensemble de ces *classes agents* ;
3. les classes contenant que des humains : elles correspondent aux comportements qui n'ont pas été reproduits par les agents. Il s'agit donc soit de manques dans le modèle agent, soit d'un ensemble d'agents trop petit dans l'espace des paramètres. Nous notons $\mathbb{C}_H$ l'ensemble de ces *classes humains*.

Nous proposons de calculer trois scores d'après ces classes. Soit $|\mathbb{C}|$ la taille d'un ensemble $\mathbb{C}$. Nous définissons :

1. $S_c = \frac{|\mathbb{C}_M|}{|\mathbb{C}_M|+|\mathbb{C}_H|}$ le score des capacités ;
2. $S_m = \frac{|\mathbb{C}_H|}{|\mathbb{C}_M|+|\mathbb{C}_H|}$ le score des manques ;
3. $S_e = \frac{|\mathbb{C}_H|}{|\mathbb{C}_M|+|\mathbb{C}_A|}$ le score des erreurs.

Seuls, ces scores ne sont pas suffisants pour calibrer la simulation multi-agent. Par exemple, si un seul agent appartient à une classe contenant beaucoup d'humains, cela signifie que ce comportement ne sera reproduit que par une petite proportion des agents alors qu'il est fréquent chez les participants humains. Dans la prochaine sous-section, nous étudions donc les effectifs des classes pour détecter les cas de sous-représentation et de sur-représentation des agents.

3.2 Calibration

Dans une simulation multi-agent, la calibration du modèle implique de trouver des valeurs de paramètres permettant d'atteindre un but global ou d'obtenir un comportement spécifique [9, 10] (voir par exemple [18]). La calibration des simulations multi-agents est un problème difficilement résolu par les méthodes de calibration classiques [9] en raison du trop grand espace de paramètres, d'un temps trop élevé de simulation et des incertitudes dans la conception du modèle. Dans les simulations participatives où les agents interagissent avec des humains, une contrainte supplémentaire est que les comportements individuels soient crédibles.

Dans notre cas, nous proposons d'utiliser les traces de comportement des humains collectées pendant la phase d'évaluation du modèle : ils définissent un ensemble de comportements valides. En d'autres termes, notre but est de diminuer les valeurs de S_e et de S_m, au profit de S_c. Si l'on considère le modèle agent comme une boîte noire qui peut produire différents comportements en fonction de ses paramètres, le processus de calibration doit garantir que : 1) le comportement de chaque agent a est crédible, donc que son ensemble de paramètres $P_a = \{p_1, \ldots, p_i\}$ (avec i le nombre de paramètres du modèle) est individuellement valide ; 2) la population reproduit globalement les mêmes comportements que les humains, dans des proportions équivalentes, d'où une distribution d'ensembles de paramètres $\mathcal{P} = \{P_1, \ldots, P_n\}$ à définir, avec n le nombre d'agents.

Concrètement, les agents membres de classes de type « erreurs » ont exhibé des comportements qui ne sont pas présents chez les humains. Leurs ensembles de paramètres P_a doivent être retirés de la liste des ensembles valides. Inversement, les manques (*i.e.* les comportements exhibés uniquement par des humains) peuvent provenir de paramètres mal choisis, en supposant que le modèle agent a effectivement la capacité de produire ces comportements.

Pour obtenir une bonne proportion de comportements valides, nous prenons en compte l'appartenance des agents et des humains aux classes.

Nous pouvons donc raffiner les ensembles précédents en :

- erreurs comportementales, lorsque aucun humain n'appartient à la classe ($H(C) = 0$),
- capacités à reproduire des comportements humains dans des classes mixtes ($\mathbb{C}_M$), séparées en trois sous-ensembles :
  - $\mathbb{C}_{sur}$ l'ensemble des classes où les agents sont sur-représentés ($A(C) \gg H(C)$),
  - $\mathbb{C}_{valide}$ l'ensemble des classes où la représentation des agents est correcte, *i.e.* où la proportion d'humains et d'agents est comparable ($A(C) \approx H(C)$),
  - $\mathbb{C}_{sous}$ l'ensemble des classes où les agents sont sous-représentés ($A(C) \ll H(C)$).
- manques, lorsque aucun agent n'appartient à la classe ($A(C) = 0$).

Après une première calibration de l'ensemble des paramètres $\mathcal{P}$, les scores ne dépendent plus du nombre d'agents mais de la proportion dans la population totale. Nous avons souvent plus d'agents que d'humains (pour des raisons expérimentales liées au temps de collecte des données). De plus, l'opérateur $\approx$ qui permet de caractériser une représentation valide nécessite de définir une limite. Nous proposons de nous baser sur la taille de la classe, *e.g.* $\delta(C) = \frac{5}{100}|C|$.

Pour améliorer la calibration, nous établissons trois scores à partir des classes mixtes $\mathbb{C}_M$, en complément des scores de manque et d'erreur :

- sur-représentation $S_{sur} = \frac{|\mathbb{C}_{sur}|}{|\mathbb{C}_M|}$ avec :

  $\mathbb{C}_{sur} = \{C_M, A(C_M) > H(C_M| + \delta|C_M|\}$

- sous-représentation $S_{sous} = \frac{|\mathbb{C}_{sous}|}{|\mathbb{C}_M|}$ avec :

  $\mathbb{C}_{sous} = \{C_M, A(C_M) < H(C_M) - \delta|C_M|\}$

- représentation valide $S_{valide} = \frac{|\mathbb{C}_{valide}|}{|\mathbb{C}_M|}$ avec :

$$\mathbb{C}_{valide} = \mathbb{C}_M \setminus (\mathbb{C}_{sur} \cup \mathbb{C}_{sous})$$

Ces scores permettent d'évaluer la qualité de la calibration courante et éventuellement de la valider. Si ça n'est pas le cas, ils permettent de modifier la proportion d'agents associés à chaque ensemble de paramètres $\mathcal{P}_i$, d'explorer de nouveaux ensembles de paramètres et de retirer ceux qui sont invalides.

3.3 Définition d'un ensemble de paramètres

Soit $\mathcal{P}_v$ l'ensemble des jeux de paramètres valides correspondant aux comportements valides $\mathcal{B}_v \subset \mathcal{B}$ ($\mathcal{B}$ désignant l'ensemble des comportements). Nous notons $simul(P_i) = b$ le comportement b produit par un ensemble de paramètres P_i et $p(b)$ la proportion d'humains exhibant ce comportement.

Puisque plusieurs jeux de paramètres peuvent produire le même comportement, la définition d'un ensemble de paramètres $P(a_i)$ avec $i \in \{1, \ldots, n\}$ pour n agents nécessite de choisir entre différents jeux de paramètres. Nous proposons de les choisir de la manière suivante : $P(a_i) = P_i \in \mathcal{P}_v$ avec une probabilité $p(P_i)$, dépendant de la proportion des comportements observés b et du nombre nb de jeux de paramètres tels que $simul(P_i) = b \in \mathcal{B}_v$ conduisent à b :

$$p(P_i) = \frac{p(b)}{nb}$$

De cette manière, les comportements qui étaient sous-représentés ont une probabilité plus élevée d'être sélectionnés, alors que les comportements sur-représentés seront moins sélectionnés.

Il est aussi possible de choisir arbitrairement l'un des ensembles de paramètres $P_i \in \mathcal{P}_v | simul(P_i) = b \in \mathcal{B}_v$ et de générer $n \cdot p(b)$ agents avec cet ensemble. Cependant, selon le type de simulation, maintenir une hétérogénéité contrôlée des agents [1] permet de produire des simulations plus réalistes.

Notons aussi que puisque nous nous limitons pour la génération des nouveaux jeux de paramètres à $\mathcal{P}_v$ (et non $\mathcal{P}$), tous les ensembles de paramètres invalides sont supprimés.

3.4 Exploration de l'espace des paramètres

Le nombre d'agents nécessaires pour couvrir tous les ensembles de paramètres possibles peut être très élevé. La première génération de jeux de paramètres s'appuie donc sur des connaissances du domaine ou sur des valeurs par défaut définies dans le modèle. En conséquence, il n'est généralement pas possible de produire tous les comportements dès le premier cycle de notre méthode d'évaluation de simulations multi-agents. Si des manques sont détectés, nous proposons alors d'utiliser une fonction d'exploration des ensembles de paramètres non

1. L'hétérogénéité est contrôlée par le fait que tous les ensembles de paramètres menant à des comportements appartenant à la même classe vont produire des comportements similaires.

encore testés :

$$P(a_i) = \begin{cases} P_i \in \mathcal{P}_v & si\ p > \gamma \\ P_k \notin \mathcal{P} & sinon \end{cases}$$

Le paramètre d'exploration γ permet de rechercher des nouveaux comportements de manière itérative, et p est une valeur aléatoire uniforme. Afin de ne pas essayer plusieurs fois des ensembles de paramètres invalides, P_k ne doit jamais avoir été sélectionné précédemment. Si P_k conduit à un comportement valide, il est ajouté à $\mathcal{P}_v$, sinon il est écarté pour la prochaine itération.

3.5 Itération de la méthode

Si tous les comportements cibles (déterminés par les classes de traces des humains) sont reproduits, un seul cycle de calibration est nécessaire. Lorsqu'il manque des comportements, l'exploration de l'espace des paramètres permet de découvrir de nouvelles classes de comportements reproduits par les agents.

Les manques et les erreurs peuvent aussi être traités par des modifications dans le modèle agent. Dans ce cas, les informations de l'étape d'annotation de la méthode d'évaluation permettent de mettre en évidence des informations sémantiques sur les comportements manquants ou erronés. Dans tous les cas, l'annotation et l'expérimentation avec des humains n'est requise qu'une seule fois, quel que soit le nombre d'itérations de la méthode de calibration des paramètres, puisque la méthode de traitement des données s'appuie sur l'agrégation des agents sur les classes de traces des humains. Ainsi, les données des agents ne modifient pas le modèle de référence constitué par les classes d'humains.

3.6 Taux de confiance

Les algorithmes de classification non supervisée peuvent produire des erreurs de classification parmi les traces humaines. En conséquence, les classes singletons peuvent représenter des comportements particuliers ou des erreurs de classification. De la même manière, l'agrégation d'agents aux classes d'humains peut, notamment à cause de l'effet de seuil, agréger un agent à une classe de comportement qui n'est pas *similaire*, ou réciproquement ne pas l'inclure dans une classe de comportement qui est *similaire*. Dans le but de vérifier cela, nous proposons de calculer un taux de confiance sur les classes résultantes.

Le taux de confiance de chaque classe dépend du nombre d'humains et d'agents dans toute la classification et dans la classe en question :
$t_A(C_M) = \frac{A(C_M)}{A(\mathbb{C})} \in [0,1]$
et $t_H(C_M) = \frac{H(C_M)}{H(\mathbb{C})} \in [0,1]$

Pour les classes mixtes, le taux de confiance est haut quand le taux de participants humains est approximativement égal à celui des agents simulés :

$$t_{conf}(C_M) = 1 - |t_A(C) - t_H(C)|$$

Par exemple, une classe contenant 9 participants sur 12 au total et seulement 2 agents sur 20 au total a un taux de $1 - \left|\frac{9}{12} - \frac{2}{20}\right| = 0{,}35$. Cette classe est alors considérée comme une classe mixte mais avec un taux de confiance bas. *A contrario*, une classe contenant 2 participants sur les 12 et 4 agents sur les 20 a un taux de $1 - \left|\frac{2}{12} - \frac{4}{20}\right| = 0{,}97$. Il y a donc une forte confiance que cette classe mixte soit réellement une capacité du modèle agent à reproduire le comportement humain adopté.

À l'inverse, pour les *classes agents*, le taux de confiance $t_{conf}(C_A)$ dépend uniquement du nombre d'agents relativement au nombre moyen d'agents $E_A(\mathbb{C}) = \frac{A(\mathbb{C})}{|\mathbb{C}_M|+|\mathbb{C}_A|}$ (respectivement $E_H(\mathbb{C}) = \frac{H(\mathbb{C})}{|\mathbb{C}_M|+|\mathbb{C}_H|}$ pour le taux de confiance $t_{conf}(C_H)$ des *classes humains*) :

$$t_{conf}(C_A) = \frac{A(C)}{E_A(\mathbb{C})}$$

$$t_{conf}(C_H) = \frac{H(C)}{E_H(\mathbb{C})}$$

Un taux de confiance élevé en une *classe d'humains* (resp. une *classe d'agents*) signifie que cette classe peut être considérée comme un manque (resp. une erreur) dans le modèle agent avec confiance.

4 Évaluation

Dans un premier temps, cette section rappelle les résultats de classification issus de [4] sur une expérimentation étudiant les comportements en simulation de conduite, puis elle les complète avec les résultats de notre extension fournissant des scores d'évaluation pour la calibration.

Le simulateur de trafic routier *ARCHISIM* [7] de l'*IFSTTAR* est évalué sur la situation spécifique suivante : sur une route bidirectionnelle, l'acteur principal rencontre un véhicule à vitesse réduite sur la file de droite et plusieurs véhicules venant en face sur la file de gauche. Le but est d'évaluer la crédibilité des comportements de conduite des agents dans le but de calibrer des simulations valides.

Pour la partie objective, nous recueillons les traces des acteurs principaux (les participants et les agents) pendant la simulation. Les indicateurs choisis par des experts du domaine sont calculés à partir de ces traces de simulation (*e.g.* la distance inter-véhiculaire ou le nombre de changements de voie). Ces indicateurs sont utilisés pour classifier les participants, et agréger les agents dans les classes. Pour la partie subjective, le questionnaire d'annotations comportemental fournit 6 sous-échelles [16] : *inexpérience*, *inattention*, *erreur de jugement*, *violation accidentelle* et *délibérée*, *risque d'accident*. 6 annotateurs ont rempli ce questionnaire à la fois pour les participants et les agents. La même méthode de traitement que celle utilisée sur les traces de simulation est appliquée sur ces scores.

22 conducteurs réguliers ont effectué l'expérimentation. Les véhicules simulés de l'*EV* étant autonomes, les situations peuvent différer. Les paramètres du modèle *ARCHISIM* sont :

- $v \in \mathbb{N}$ la vitesse désirée (en km/h),
- $reglmnt \in [0{,}100]$ la volonté de conduire selon le code de la route,
- $infra \in [0{,}100]$ la capacité de contrôle du véhicule selon l'infrastructure,
- $trafic \in [0{,}100]$ la flexibilité du temps inter-véhiculaire selon le trafic,
- $soupl \in [0{,}100]$ l'agressivité du conducteur (ne tenant pas compte des courtes variations),
- $exp \in \{0{,}1\}$ l'expérience du conducteur.

Dans cette itération, les paramètres utilisés sont les valeurs moyennes pour chaque paramètre, tandis qu'un paramètre est modifié pour chaque agent. Le paramètre v est choisi parmi $\{100, 110, 120, 130, 140\}$; les quatre paramètres suivant ($reglemnt, infra, trafic, soupl$) sont par défaut à 50, et à 25 ou 75 sinon ; enfin exp est à 0 ou à 1.

4.1 Résultats de classification

Les classifications de traces et d'annotations sont comparées comme illustré dans la figure 2. Il y a 2 classes de comportement issues des traces : 1) *Logs cluster 1* est composée des acteurs principaux qui n'ont pas essayé de dépasser le véhicule à vitesse réduite. Étant une classe mixte, c'est une capacité du modèle agent à reproduire un comportement humain qui est celui de choisir de ne pas dépasser. 2) *Logs cluster 2* contient des acteurs principaux qui ont dépassé le véhicule à vitesse réduite. Comme cette classe est uniquement composée de participants, c'est donc un manque dans le modèle agent : les agents ne peuvent pas choisir de ne pas dépasser comme les humains le font.

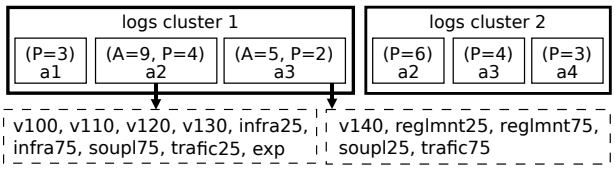

FIGURE 2 – Comparaison des acteurs principaux entre la classification de traces (avec les participants P et les agents A), et la classification d'annotations regroupés par classe avec leur numéro ($a\#$), en fonction de leurs paramètres.

Il y a 4 classes de comportements issues des annotations : 1) la classe d'annotations 1 a été considérée par les annotateurs comme ayant le style de conduite le plus dangereux, aucun agent n'a été considéré dangereux. 2) la classe d'annotations 2 a été annotée comme correspondant à des conducteurs prudents. Étant une classe mixte, le comportement normatif humain peut donc être considéré comme partiellement reproduit. 3) la classe d'annotations 3, mixte également, est une petite classe d'annotations dont les acteurs principaux ont été jugés comme étant des conducteurs ordinaires. 4) la classe d'annotations 4 contient des acteurs principaux considérés comme légèrement dangereux, comportement non reproduit par les agents.

Dans la classe *logs cluster 1*, les indicateurs n'ont pas permis de distinguer la classe d'annotations 1 du reste des acteurs principaux, tandis que ces acteurs principaux ont essayé de dépasser en vain. Pareillement, les participants de la classe d'annotations 4 n'ont pas été séparés de la classe *logs cluster 2* en une nouvelle classe.

4.2 Scores

Les scores correspondant à la classification des annotations dans cette expérimentation sont les suivants :

- le score d'erreur est $S_e = 0$ (car $|\mathbb{C}_A| = 0$),

- le score de manque est :
  $S_m = \frac{|\mathbb{C}_H|}{|\mathbb{C}_M|+|\mathbb{C}_H|} = \frac{1}{1+1} = \frac{1}{2}$.
- le score de capacité est :
  $S_c = \frac{|\mathbb{C}_M|}{|\mathbb{C}_M|+|\mathbb{C}_H|} = \frac{1}{1+1} = \frac{1}{2}$.

Le taux de confiance en la classe 1 comme étant une capacité du modèle est :

$$t_{conf}(C_1) = 1 - \left|\frac{14}{14} - \frac{9}{22}\right| \approx 0{,}41$$

$t_{conf}(C_1)$ n'étant pas proche de 0, cette classe peut donc être considérée avec confiance comme étant une capacité du modèle.

La classe de traces 2 est une classe composée uniquement d'humains, le taux de confiance est alors calculé de la manière suivante :

$$E_H(\mathbb{C}) = \frac{22}{1+1} = 11$$

$$t_{conf}(C_2) = \frac{13}{11} \approx 1{,}18$$

$t_{conf}(C_2)$ est proche de 1 signifiant que cette classe composée uniquement d'humains peut être traitée avec confiance comme étant un comportement manquant réel dans le modèle agent.

Ces scores permettent de trouver les nombres de comportements adéquats et inadéquats, et donc de produire l'ensemble des jeux de paramètres qui sont valides : $P \in \mathcal{P}_v = \{v100,\ v110,\ v120,\ v130,\ infra25,\ infra75,\ soupl75, trafic25, exp\} \cup \{v140, reglmnt25,\ reglmnt75,\ soupl25, trafic75\}$.

Il n'y a qu'une classe mixte. Cette dernière est sur-représentée, en effet pour un $\delta = \frac{5}{100}$ (*i.e.* peu de tolérance aux variations de proportions entre les comportements humains et agents) :

$$A(C_1) > H(C_1) + \delta|C_1| \Leftrightarrow 14 > 9 + \frac{5}{100} \times 23$$

Les scores de représentativité correspondants sont : $S_{sur} = \frac{|\mathbb{C}_{sur}|}{|\mathbb{C}_M|} = 1$, $S_{sous} = \frac{|\mathbb{C}_{sous}|}{|\mathbb{C}_M|} = 0$, et $S_{valide} = \frac{|\mathbb{C}_{valide}|}{|\mathbb{C}_M|} = 0$.

Les scores de type de classe ont permis de quantifier les erreurs (0), capacités (0,5), et manques (0,5). Les taux de confiance ont permis de s'assurer que la classe mixte est bien une capacité du modèle d'agent ($t_{conf}(C_1) \approx 0{,}41$) et que la classe d'humains est bien un comportement humain manquant ($t_{conf}(C_2) \approx 1{,}18$). Les scores de représentativité montrent que la calibration originale sur-représente le comportement humain (correspondant à la classe mixte) dans la population d'agents.

Une fois ces scores calculés, il est possible de calibrer une nouvelle population d'agents.

4.3 Calibration

Nous avons trouvé que seule une partie des comportements humains étaient reproduits. Dans ce cas, il existe deux possibilités pour l'utilisateur de la simulation : soit entrer dans un cycle d'exploration de l'espace des paramètres, soit calibrer le système en utilisant la première analyse. Dans cette partie, nous nous concentrons sur la seconde possibilité.

Nous avons vu que tous les jeux de paramètres étaient valides mais sur-représentés, et qu'il y avait plusieurs comportements manquants. Par conséquent, afin d'obtenir des agents reproduisant le comportement humain, nous calibrons les nouveaux ensembles de paramètres des agents en les choisissant parmi ceux valides dans la première expérimentation et en raffinant les proportions grâce aux classes d'annotations.

La première classe contient 9 traces d'agents et 4 traces de participants, tandis que la seconde classe contient 5 traces d'agents et 2 traces de participants. Nous obtenons donc, en calculant leur probabilité d'apparition dans la prochaine calibration : $p(v100) = p(v110) = p(v120) = p(v130) = p(infra25) = p(infra75) = p(soupl75) = p(trafic25) = p(exp) = \frac{4/6}{9}$ et $p(v140) = p(reglmnt25) = p(reglmnt75) = p(soupl25) = p(trafic75) = \frac{2/6}{5}$.

De cette manière, les agents représentent stochastiquement la densité de comportements humains dans la simulation. Il est à noter que cette nouvelle calibration ne change pas les scores de capacités et de manques, qui sont basés sur la proportion de comportements humains correctement reproduits. Cependant, il réduit le score d'erreur à 0 en n'utilisant que des ensembles de paramètres valides, et améliore le score de représentativité en choisissant des ensembles de paramètres pour les agents selon les comportements observés chez les humains.

Discussion

Notre méthode introduit des métriques pour mesurer les scores résultant et pour corriger les paramètres des agents selon une telle étude. Les scores d'erreurs, de manques, et de capacités permettent au concepteur de la simulation multi-agent de trouver combien d'archétypes de comportement humain ont été correctement reproduits, combien de comportements agents ne devraient pas apparaître, et combien de comportements humains sont manquants. Le taux de confiance donne des indications sur la fiabilité des classes en fonction de leurs effectifs. Ensuite, en étudiant uniquement les archétypes des comportements correctement reproduits, les scores de calibration donnent de l'information sur les proportions de chaque comportement et leurs relations à la calibration des agents.

Une des originalités de ce processus de calibration est le contexte des simulations participatives. La fonction cible du processus de calibration n'est pas comme habituellement [9] au niveau macroscopique mais à un niveau individuel. La réalité virtuelle requière que chaque agent adopte des comportements crédibles, *i.e.* des comportements pouvant être produits par des humains. Dans ce contexte, nous retirons premièrement les ensembles de paramètres qui ne produisent pas de comportements valides. Nous calibrons ensuite les proportions d'agents avec les ensembles de paramètres restants selon les données des participants humains. Un unique cycle de notre méthode assure que les comportements valides sont détectés et qu'ils sont produits en des proportions correctes, nonobstant les comportements manquants.

Dans le cas d'une boîte noire où le modèle agent est inconnu et ne peut être modifié, si des comportements sont manquants alors une solution est d'explorer l'espace des paramètres afin de trouver de nouveaux comportements d'agents. Ces nouvelles étapes ne requièrent pas une autre expérimentation avec des participants humains, puisque les données de référence sont déjà disponibles. Chaque nouveau cycle permet de trouver potentiellement de nouveaux ensembles de paramètres, soit dans des classes déjà *mixtes*, soit dans les classes de *manques* précédentes. Cela permet aussi de trouver quelles zones de l'espace des paramètres produisent des comportements invalides.

Dans le cas d'une boîte blanche où le modèle agent est connu et peut être potentiellement modifié, les données d'annotations expliquent les comportements manquants et les comportements erronés, permettant ainsi d'améliorer le modèle agent [10]. De plus, l'exploration des ensembles de paramètres peut être guidé par la connaissance du modèle [9].

5 Conclusion & perspectives

Cet article présente une méthode semi-automatique de calibration de simulations multi-agents participative basée sur la combinaison de classifications non supervisées de traces de simulation et d'annotations par des participants. Ces classifications portent à la fois sur les comportements des agents et des participants, comparés dans la même situation. L'expérimentation permet de définir un ensemble initial de traces valides qui servent de points de référence pour la calibration du modèle multi-agent. La calibration des paramètres du modèle suit une approche itérative. À chaque itération, nous obtenons les manques et les erreurs du modèle afin de raffiner l'espace des paramètres. Nous générons de nouveaux agents qui sont agrégés aux classes précédentes et nous calculons de nouveaux scores pour l'itération suivante. Notre méthode de validation a été appliquée à la simulation de trafic routier et a montré que des paramètres peuvent être correctement associés à des catégories de comportement.

L'originalité de cette approche est double. Premièrement, elle combine une analyse automatique des comportements des agents via les traces de simulation avec une analyse subjective basée sur l'évaluation humaine des comportements des agents, dans le but de définir un contexte spécifique à la situation. Cette combinaison permet une analyse de quel espace des paramètres des agents virtuels produit quel comportement perçu. Secondement, elle itère l'analyse de la classification afin de raffiner les capacités des agents à reproduire des comportements humains, tout en réduisant les manques et en supprimant les erreurs.

Plusieurs extensions doivent être examinées. Premièrement, la méthode d'agrégation dépend d'un taux de tolérance dont la valeur peut impacter la qualité des résultats : cet impact devra être vérifié. Secondement, la convergence du modèle n'a pas encore été étudiée. Notre algorithme de classification donne des scores qui pourraient être utilisés afin de stopper le processus, mais une preuve de convergence est nécessaire lorsque le cycle n'est pas utilisé pour ex-

plorer de nouveaux paramètres.

Références

[1] T. Bosse, M. Hoogendoorn, M. C. Klein, J. Treur, and C. N. Van Der Wal. Agent-based analysis of patterns in crowd behaviour involving contagion of mental states. In *Modern Approaches in Applied Intelligence*, pages 566–577. Springer, 2011.

[2] P. Caillou and J. Gil-Quijano. Simanalyzer : Automated description of groups dynamics in agent-based simulations. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pages 1353–1354, 2012.

[3] S. Campano, N. Sabouret, E. de Sevin, and V. Corruble. An evaluation of the core computational model for affective behaviors. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 745–752, 2013.

[4] K. Darty, J. Saunier, and N. Sabouret. Agents behavior semi-automatic analysis through their comparison to human behavior clustering. In *Intelligent Virtual Agents*, pages 154–163. Springer, 2014.

[5] K. Darty, J. Saunier, and N. Sabouret. Analyse des comportements agents par agrégation aux comportements humains. In *22èmes Journées Francophones sur les Systèmes Multi-Agents (JFSMA 2014)*. Cépaduès, 2014.

[6] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony : A survey of evaluation methods across disciplines. *Affective Computing, IEEE Transactions on*, 3(3) :349–365, 2012.

[7] A. Doniec, R. Mandiau, S. Piechowiak, and S. Espié. A behavioral multi-agent model for road traffic simulation. *Engineering Applications of Artificial Intelligence*, 21(8) :1443–1454, 2008.

[8] A. Drogoul, B. Corbara, and D. Fresneau. Manta : New experimental results on the emergence of (artificial) ant societies. *Artificial Societies : the computer simulation of social life*, pages 190–211, 1995.

[9] M. Fehler, F. Klügl, and F. Puppe. Techniques for analysis and calibration of multi-agent simulations. In *Engineering Societies in the Agents World V*, pages 305–321. Springer, 2005.

[10] M. Fehler, F. Klügl, and F. Puppe. Approaches for resolving the dilemma between model structure refinement and parameter calibration in agent-based simulations. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 120–122. ACM, 2006.

[11] J. Gonçalves and R. J. F. Rossetti. Extending sumo to support tailored driving styles. *1st SUMO User Conference, DLR, Berlin - Adlershof, Germany*, 21 :205–211, 2013.

[12] B. Lacroix, P. Mathieu, and A. Kemeny. Formalizing the construction of populations in multi-agent simulations. *Engineering Applications of Artificial Intelligence*, 2012.

[13] J. C. Lester, S. A. Converse, et al. The persona effect : affective impact of animated pedagogical agents. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 359–366. ACM, 1997.

[14] P. Mathieu and O. Brandouy. A generic architecture for realistic simulations of complex financial dynamics. In *Advances in Practical Applications of Agents and Multiagent Systems*, pages 185–197. Springer Berlin Heidelberg, 2010.

[15] C. Pelachaud. Modelling multimodal expression of emotion in a virtual agent. *Phil. Trans. R. Soc. B : Biological Sciences*, 364(1535) :3539–3548, 2009.

[16] J. Reason, A. Manstead, S. Stradling, J. Baxter, and K. Campbell. Errors and violations on the roads : a real distinction ? *Ergonomics*, 33(10-11) :1315–1332, 1990.

[17] P. Taillandier and A. Drogoul. Supervised feature evaluation by consistency analysis : application to measure sets used to characterise geographic objects. In *Knowledge and Systems Engineering (KSE), 2010 Second International Conference on*, pages 63–68. IEEE, 2010.

[18] A. Veremme, É. Lefevre, G. Morvan, D. Dupont, and D. Jolly. Evidential calibration process of multi-agent based system : An application to forensic entomology. *Expert Systems with Applications*, 39(3) :2361–2374, 2012.