7ème Symposium MaDICS 2025 - 2ème année de thèse

Jumeau numérique de processus métier : approche probabiliste augmentée par GNN

Guilhem Nespoulous ⋆,†, Frédéric Bertrand †, Myriam Maumy‡, Yoann Valéro ⋆

⋆ QAD Process Intelligence, France † Université de Technologie de Troyes, France, laboratoire LIST3N ‡ EHESP

Introduction & définitions

Le ***process minining*** est un domaine scientifique basé sur les données provenant de processus métiers. Ces données sont caractérisées par un journal d'événement ou **envent log** qui contient un ensemble d'**événements**. Chaque événement est composé d'un triplet contenant un **identifiant d'unité**, une **activité** et un **horodatage**. Les événements contiennent également parfois des observations de certaines covariables.

L'ensemble des événements d'une unique unité est appelé un **parcours** d'unité. La séquence des activités du parcours d'une unité est appelée une **trace**.

ID d'unité	Activité	horodatage
1	Réception CV	11/08/24 11:23:47
1	Entretien RH	22/08/24 19:44:28
2	Réception CV	23/08/24 08:11:02
2	Entretien Technique	26/08/24 17:32:36
1	Refus	29/08/24 13:56:33
2	Embauche	30/08/24 15:02:01
...	...	...

Fig1 Exemple de journal d'événements.

Un **jumeau numérique** est une réplique d'une entité physique. Il permet d'effectuer des simulations multidimensionnelles à différentes échelles.

Ces simulations, peuvent permettre d'étudier les effets de potentielles modifications du jumeau physique qui seraient impossibles (ou trop coûteuses) à implémenter. Cette étude se fait en analysant la réponse du jumeau numérique qui aura été mis à jour en conséquence, on parle de ***what-if scenario***.

Dans le cadre du *process mining*, les scénarios que l'on peut chercher à évaluer peuvent comprendre une combinaison des éléments suivants :

- Modification du flux d'entrée d'unités
- Changement de répartition des traces
- Changement de certaines covariables.
- Etc.

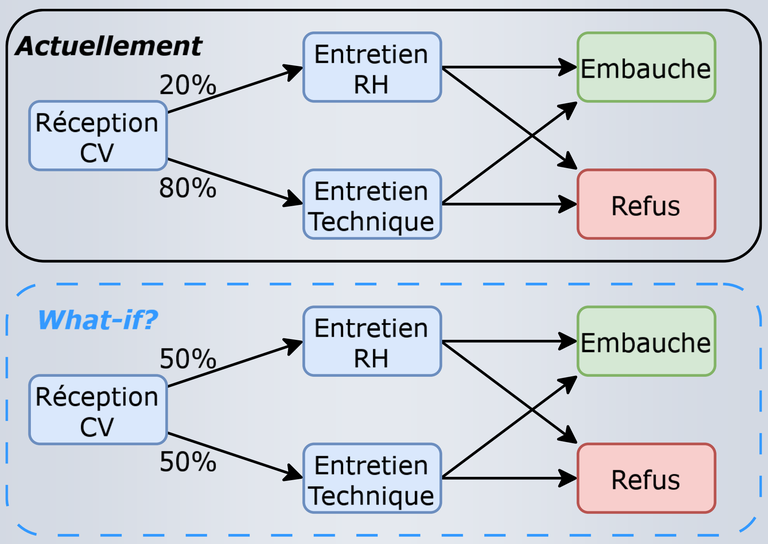


Fig.2 Exemple de *what-if scenario*.

Objectif

L'objectif de cette thèse est de définir une méthodologie de mise en place de jumeau numérique de processus métier qui satisfait les points suivants :

- Construction du jumeau avec des **paramètres explicites et intelligibles pour les utilisateurs**
- Prise en compte et restitution de l'**incertitude**
- Performance sur une grande **variété de processus**
- **Explicabilité**, implémentation du principe de « l'**écart au nominal** »
- Simulation à **multiples échelles** (unité, processus)
- Choix libre des variables à évaluer (indicateurs liés aux traces, dernières activités des parcours, temporalité,)

Comme le souligne également W.V. van der Aalst dans Business Process Simulation Revisited [1], les données issues des journaux d'événements offrent une vision limitée du système global sous-jacent au processus. Elles ne permettent généralement d'expliquer qu'une fraction de la variance observée dans les comportements des processus.

Nous pensons qu'une **approche de simulation stochastique est essentielle** pour prétendre à un véritable intérêt industriel [2].
Notre objectif n'est donc pas de déterminer l'impact sur une variable d'un scénario sur *le déroulé le plus probable*. Mais plutôt d'estimer l'impact du scénario sur la fonction de densité estimée de la variable cible sans conditionner cela à un déroulé spécifique.

Méthodologie

La génération nominale est effectuée à l'aide d'un modèle graphique probabiliste à effet aléatoires.

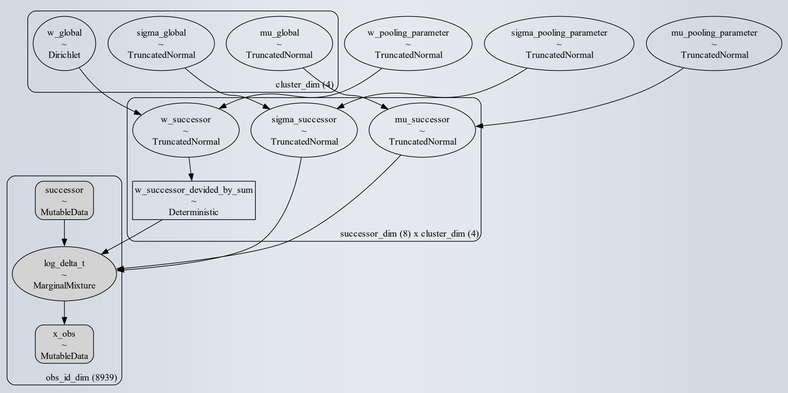


Fig.3 : Exemple de modèles bayésien hiérarchique simplifié permettant d'échantillonner l'intervalle de temps avant le prochain événement.

Nous pénalisons les paramètres d'effets en fonction de la complexité de l'interaction. Soit d le nombre de variables qui régissent l'effet e. Pour paramétriser un apriori de Laplace nous choisissons par exemple :

$$e \sim Laplace(0, \frac{1}{d})$$

De plus, afin de fluidifier les simulations en réduisant le nombre de paramètres nous proposons deux options :

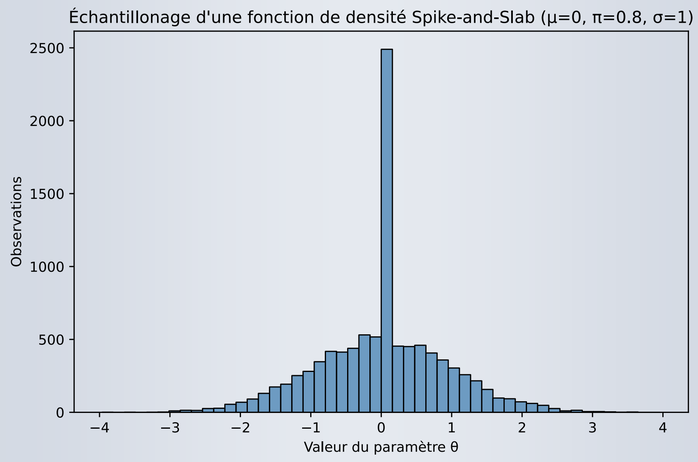
1. option ante hoc où les a priori sont de type « *spike and slab* ».
2. option post hoc, l'a priori utilisé est une loi de Laplace. Un seuil est utilisé pour neutraliser les effets négligeables.

Un apriori *spike and slab* avec une slab gaussienne peut être obtenu comme suit pour un paramètre θ:

$$\theta \sim (1 - z_i) \cdot \delta_0 + z_i \cdot \mathrm{N}(0, \sigma^2)$$

Avec :

- $z_i \sim Bernoulli(\pi)$
- δ_0 fonction de dirac en 0 qui correspond au *spike*
- $N(0, \sigma^2)$ la *slab* gaussienne
- π est la probabilité de prise en compte de l'effet



Un réseau de neurones en graphe (GNN) est utilisé afin d'ajuster les paramètres nominaux en capturant des interactions complexes entre les variables explicatives, ainsi que des dépendances temporelles portant sur plusieurs événements précédant celui à prédire. Contrairement aux approches purement markoviennes, les GNN permettent de modéliser une part de la variance du processus qui résulte de relations structurelles à plus long terme.
Pour cela, nous transformons nos données en un ensemble de multigraphes dirigés où chaque multigraphe correspond au parcours tronqué d'une unité [5].

Etapes de l'étude d'impact d'un *what-if scenario* :

1. Échantillonnage a posteriori des paramètres, phase de calibration. Réalisé à l'aide d'un algorithme MCMC (NUTS, GIBS etc.) pour atteindre la convergence.
2. Échantillonnage post calibration des paramètres. Implémentation d'un *what-if scénario* (optionnel) encodé sous forme d'une modification (ou d'un ajustement) des paramètres du modèle, afin de simuler l'effet d'une politique, d'une contrainte ou d'un changement de condition.
3. Définition des variables cibles ou KPIs spécifiées par l'utilisateur (durée, coûts, variants, etc.). Une fonction est générée pour extraire automatiquement ces KPIs à partir des simulations réalisées.
4. Pour chaque simulation, un jeu de paramètres est tiré aléatoirement parmi ceux obtenus lors de l'échantillonnage. Le processus est simulé avec ce jeu de paramètres, les KPIs sont calculés et enregistrés. Le GNN corrige les paramètres nominaux.
5. Une analyse statistique est ensuite effectuée sur l'ensemble des KPIs (moyenne, médiane, intervalles de confiance...).

Premiers résultats et perspectives

Mesurer la qualité d'une méthodologie présentée est une tâche complexe. Il y a deux approches différentes :

- Evaluation ponctuelle de la distance [4] entre les trajectoires simulées et les données observées (par validation croisée). Une telle approche ignore la structure probabiliste du processus simulé et tend à favoriser les prédictions moyennes, au détriment de la représentation de la variance et des trajectoires moins probables mais cohérentes.
- Evaluation sur données synthétiques. Cette approche nécessite d'émettre des hypothèses fortes sur les caractéristiques empiriques des processus métier afin de pouvoir générer des processus « représentatifs ». Pour autant, elle permet une comparaison par méthodes de distance de Wasserstein.

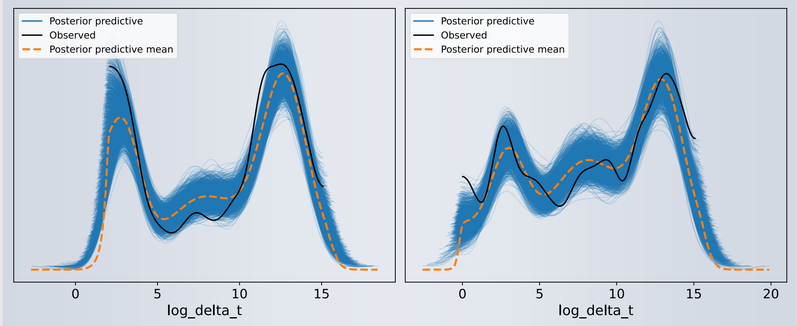


Fig.5 Échantillonnage a posteriori des intervalles de temps conditionnés à l'activité du successeur, pour deux activités différentes. Données provenant du jeu de données Helpdesk[3].

Outils utilisés

Pour la conception et l'entraînement des réseaux de neurones en graphe, nous avons principalement utilisé **PyTorch** et PyTorch Geometric (**PyG**), qui offrent une grande flexibilité pour la modélisation sur graphes. Concernant les modèles probabilistes, nous avons eu recours à **Pyro**, **Stan** et **PyMC**.

Références

[1] Van der Aalst, W.M.P. (2010). Business Process Simulation Revisited. In: Barjis, J. (eds) Enterprise and Organizational Modeling and Simulation. EOMAS 2010. Lecture Notes in Business Information Processing, vol 63. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15723-3_1

[2] S. Singh, M. Weeber, and K.-P. Birke, "Advancing digital twin implementation: a toolbox for modelling and simulation," Procedia CIRP, vol. 99, pp. 567-572, 2021. DOI: https://doi.org/10.1016/j.procir.2021.03.078.

[3] Verenich, Ilya (2016), "Helpdesk", Mendeley Data, V1, doi: 10.17632/39bp3vv62t.1

[4] Chapela-Campa, D., Benchekroun, I., Baron, O., Dumas, M., Krass, D., Senderovich, A. (2023). Can I Trust My Simulation Model? Measuring the Quality of Business Process Simulation Models. In: Di Francescomarino, C., Burattin, A., Janiesch, C., Sadiq, S. (eds) Business Process Management. BPM 2023. Lecture Notes in Computer Science, vol 14159. Springer, Cham. https://doi.org/10.1007/978-3-031-41620-0_2

[5] Weinzierl, S. (2022). Exploring Gated Graph Sequence Neural Networks for Predicting Next Process Activities. In: Marrella, A., Weber, B. (eds) Business Process Management Workshops. BPM 2021. Lecture Notes in Business Information Processing, vol 436. Springer, Cham. https://doi.org/10.1007/978-3-030-94343-1_3

MEMBER OF EUT+ EUROPEAN UNIVERSITY OF TECHNOLOGY