Les données longitudinales, de trajectoire et de télémonitoring : qu’en fait-on ? Exemple des données du sommeil

Introduction

Nous cumulons de nombreuses données tout au long de notre vie, en particulier dans le domaine de la santé, concernant nos caractéristiques physiques, socioéconomiques, nos différentes maladies, traitements, hospitalisations, etc. Ce sont la base de notre trajectoire de données individuelle relatant tous nos changements et développement au cours de notre vie.

Lors d’études transversales nous ne prenons qu’un point de temps spécifique mais nous ne pouvons pas analyser les changements et trajectoires des individus au cours du temps. C’était pourtant le design d’étude le plus couramment utilisé. Une solution est donc d’évaluer les mesures de façon répétée dans le temps : ce sont les données longitudinales. Généralement, les données longitudinales sont centrées sur une période de temps précise avec des mesures prises à intervalles réguliers qui permet d’étudier par exemple, des facteurs de risques ou des effets de traitements1. Dans notre contexte de données du sommeil, l’utilisation de ces données longitudinales permet notamment d’étudier les données relatives à la PPC (Pression Positive Continue) comme la date de début du traitement du patients, ses arrêts et ses reprises du traitement au cours du temps, ses données manquantes ou extrêmes, ses changements de prestataires, ses sorties d’études, ses autres traitements, ses comorbidités, son décès.

Plus récemment, et particulièrement lors du COVID, l’arrivée du télémonitoring a permis une expansion du modèle de données longitudinales. Le télémonitoring est un suivi régulier des patients permettant de s’assurer de l’efficacité, de la tolérance et de l’observance des traitements2 (+ ref R.Tamisier 2020 ?). Ce modèle d’étude sera de plus en plus utilisé et particulièrement dans le domaine du sommeil avec l’analyse des données de PPC par exemple puisqu’il permet d’étudier l’observance et l’efficacité du traitement (IAH résiduel, fuites, …).

Ce design d’étude représente toutefois un défi pour les statisticiens et méthodologistes : la présence de données manquantes, la corrélation de certaines mesures répétées au sein d’un même cluster, la temporalité des observations : les données sont généralement ordonnées dans le temps mais les intervalles d’évaluation des mesures peuvent être inégaux, sont des contraintes que les modèles statistiques doivent prendre en compte1. De nouvelles méthodes sont développées pour analyser ces données tout en répondant au maximum à ces critères, mais comment choisir le bon modèle d’analyse statistique ?

Dans une première partie, nous ferons un tour non exhaustif des principaux modèles d’analyse dans le cadre des données du sommeil, puis nous définirons dans quel contexte utilisé ces modèles et enfin nous donnerons deux exemples d’analyse de données répétées dans le domaine du sommeil.

# I/ Les différents modèles statistiques

Avec l’expansion des études de données longitudinales, de nombreuses méthodes statistiques se sont développées essayant de répondre aux différentes caractérisations des données (types de variables, composition des variables, objectif d’étude, etc.). Dans cette partie, nous décrivons les méthodes les plus couramment utilisées pour les analyses des données du sommeil.

Tout d’abord, nous allons décrire les modèles dits de classes latentes. Un modèle de classes latentes permet d’introduire une quantité non observée ou d’inclure des variables aléatoires qui ne sont pas directement observées. Les individus sont associés à des sous-groupes de trajectoire latente selon leurs symptômes, comportements observés. C’est un modèle très flexible qui prend en compte de nombreuses caractéristiques de données telles que les données manquantes partielles, les mesures répétées discrètes ou les covariables variant dans le temps.

Puis nous aborderons des modèles de régressions mixtes, joints, de cross-corrélation et un modèle dynamique.

1. **GMM (Growth Mixture Modelling)**

*Description –* Ce modèle paramétrique est un modèle de mélange fini qui analyse des données longitudinales de suivi aussi bien continues que catégorielles. Il permet une représentation alternative des changements de trajectoire souvent modélisés via des modèles polynomiales et il est, en particulier, efficace pour représenter des trajectoires de formes complexes. Ce modèle décrit les différents changements de trajectoire longitudinal en définissant des sous-groupes de données et en définissant les changements au sein des sous-groupes. Ces sous-groupes correspondent à des classes et cette variable de classe est latente ou non observée. Plus précisément, le modèle GMM estime la trajectoire moyennes des courbes pour chaque classe et permet les variations entre les individus d’une même classe. Pour chaque trajectoire, le modèle estime les coordonnées à l’origine, la pente et la variance des données de trajectoire en maximisant la fonction de log-vraisemblance. Les coordonnées à l’origine aléatoires permettent de capturer chaque niveau initial des individus sur les données mesurées tandis que la pente aléatoire permet de capturer chaque changement individuel de ces mesures répétées par une fonction temporelle. Ensuite, lors de l’attribution des noms de chaque classe latente il est nécessaire de prendre en compte les différences uniques de chaque pattern (différentes formes de trajectoires, différents niveaux : élevés, faibles, moyens, différentes échelles). Une fois la trajectoire bien définie, elle pourra être utilisée comme variable dépendante ou non pour explorer les prédicteurs de la trajectoire de santé et explorer leur contribution aux résultats futurs en matière de santé.

Il y a cependant quelques éléments importants à prendre en considération comme le fait que la trajectoire finale dépend des paramètres de départ. Des effets aléatoires peuvent ainsi être introduit dans le modèle GMM et leur distribution ainsi que des effets fixes peuvent être spécifiés pour chaque classe. Les covariables (variables ou non dans le temps) sont acceptées pour explorer la relation entre les caractéristiques d’origine et le pattern de trajectoire mais uniquement une fois le modèle final défini. Elles ne doivent pas modifier le profil du modèle sélectionné mais influenceront toutefois le modèle final. L’ajout de prédicteurs est également autorisé pour déterminer la variable latente.

*Objectifs –* Examiner des trajectoires uniques de données longitudinales avec des mesures de suivi répétées ; identifier des patterns de trajectoire (ex : des patterns de non observance de traitement) ; identifier des classes non observées (latentes) et décrire les changements longitudinaux au sein de chacune de ces classes non observées (latentes) ;

*Procédé –* Le modèle GMM applique une procédure itérative et requiert des décisions à priori. La procédure est la suivante :

1. Hypothétiser le nombre attendu de classes latentes et la forme de la courbe de trajectoire pour chaque classe (linéaire, quadratique, cubique, …) ;
2. Caractériser le modèle : prendre des décisions concernant les paramètres de trajectoire de la courbe (coordonnées à l’origine, variance et covariance de la pente) ;
3. Estimation du modèle : par le maximum de vraisemblance ou par des méthodes Bayésiennes, les implémentations couramment utilisées sont la procédure EM (Espérance-Maximisation) ou MCMC (Markov Chain Monte Carlo) où la variable de classification des classes non observées est traitée comme une variable latente. Ces algorithmes utilisent un procédé itératif pour obtenir les estimations des paramètres et les estimations à postériori de probabilité d’appartenance de chaque individu à chaque classe possible par rapport aux données observées. Chaque individu appartiendra à la classe pour laquelle il a la plus haute probabilité d’appartenance ;
4. Sélection du modèle : choisir le modèle donnant la meilleure représentation des données observées par différents tests (Lo-Mendell Rubin adjusted likelihood ratio test (LRT, p < 0.05), Parametric bootstrapped likelihood ratio test (BLRT, p < 0.05), BIC (un BIC minimal correspond au meilleur modèle), AIC (un AIC minimal correspond au meilleur modèle), convergence et capacité à distinguer les différentes classes (entropie proche de 1.0)) et par validation des paramètres du modèle (proportion de la population dans chaque classe (> 5%), probabilité d’appartenance à la classe (proche de 1.0), parcimonie et utilité/sens des classes latentes pour l’étude).

*Avantages –* Prend en compte les données manquantes partielles et les données manquantes des outcomes sous l’hypothèse MAR (Missing At Random) ; accepte les résidus corrélés ; capture les différences inter et intra individus au cours du temps ; identifie de manière probabiliste les sous-groupes homogènes au sein d’associations hétérogènes plus importantes et représente l’hétérogénéité non observée en déduisant l’appartenance de chaque individu à des classes latentes à partir des données de trajectoire du modèle ; c’est un modèle approprié pour modéliser la stabilité et les changements au cours du temps grâce à leur capacité à examiner les différences inter individuelles dans les processus intra individuels ; permet à la trajectoire d’être qualitativement différente au cours du temps selon les différents groupes d’individus.

*Limites –* Estime beaucoup de paramètres ; son interprétation peut être complexe ; certains paramètres doivent être choisis à priori ; c’est une technique exploratoire contrainte qui recherche l’histoire que les données tentent de raconter mais il est possible que le modèle identifie de faux groupes.

Cependant des solutions peuvent être apportées à certaines limites comme la réplication des résultats sur de nouvelles données qui identifieraient les fausses classes.

*Ajouter une image illustrant le modèle ?*

1. **LTA (Latent Transition Analysis)**

*Description –* Le modèle LTA est un modèle de mélange fini, semi-paramétrique de données longitudinales qui analyse les changements des variables catégorielles, des tableaux de contingence, mesurées au cours du temps. Le modèle LTA se base donc sur la théorie des classes latentes où des sous-groupes homogènes d’individus peuvent être identifiés par leur réponse aux mesures répétées dans le temps. Il utilise des données observées d’un jeu de données catégorielles pour définir une variable latente à chaque point de temps. Cependant, lorsque la variable catégorielle contient trop de catégories, il est recommandé de réduire au maximum son nombre de catégories et il est également recommandé de n’utiliser ce modèle qu’avec un nombre de points de temps inférieur ou égal à 6. De plus, ce modèle permet à chaque individu de changer de classes au cours du temps. Ainsi, il étudie la probabilité de transition d’un individu d’une classe à un point de temps, à une autre classe au point de temps suivant. Pour cela, il étudie la matrice de probabilité de transition entre deux points de temps consécutifs. L’estimation du modèle nous donne donc des informations concernant la probabilité d’appartenance au statut latent au 1er point de temps, la proportion de la population au sein de chaque classe latente à chaque point de temps, la probabilité conditionnelle de passer d’un statut latent à un autre au cours du temps, la probabilité conditionnelle d’appartenance de réponse à l’item[[1]](#footnote-1) au statut latent et, à tout point de temps la probabilité des classes à postériori. Les paramètres du modèle sont estimés par maximisation de la fonction de vraisemblance via l’algorithme EM ou par une méthode Bayésienne via l’algorithme MCMC. Attention toutefois au fait que l’algorithme MCMC n’est pas conseillé lorsque l’échantillon est trop petit ou lorsque la probabilité de réponse à l’item est éloignée de 0 ou 1.

Pour ce modèle, les covariables sont acceptées mais le nombre de classes doit être connu avant leur ajout pour éviter le potentiel changement de nombre de classes du modèle final avec et sans covariables.

*Objectifs –* Etudier la probabilité de transition d’un individu d’une classe à un point de temps, à une autre classe au point de temps suivant.

*Procédé –* Le modèle LTA applique une procédure itérative et requiert des décisions à priori basées sur les connaissances du domaine d’étude. La procédure est la suivante :

1. Choix du nombre de classes latentes basé sur le résultat de test d’hypothèse ;
2. Caractérisation du modèle : statuer sur l’invariance temporelle de la probabilité de réponse à l’item\*, sur l’invariance des mesures pour la probabilité de transition et sur l’ajout de covariables ;
3. Estimation du modèle : le choix du modèle doit être fait avant d’appliquer les modèles. Les paramètres sont estimés par maximisation de la vraisemblance à l’aide d’un algorithme EM ou par une approche Bayésienne à l’aide d’un algorithme MCMC
4. Sélection du modèle : par AIC ou BIC.

*Avantages –* Le modèle LTA permet de modéliser un changement au cours du temps et d’identifier les prédicteurs de ces changements ; compare les différentes classes pour déterminer les effets des traitements et pour évaluer la contribution des différentes mesures pour chaque classe latente ; il est approprié pour modéliser la stabilité et le changement temporel par sa capacité à évaluer les différences inter-individu dans les processus intra-individu.

*Limites –* L’étude doit inclure un large échantillon d’individu car le modèle doit estimer de nombreux paramètres, par exemple la matrice de probabilité de transition, notamment lors de l’utilisation de l’algorithme MCMC qui utilise généralement le ‘burn-in’ pour éliminer la dépendance des valeurs de départ ou des échanges de labels des clusters; en augmentant le nombre de points de temps (> 6), le modèle LTA devient plus complexe car le modèle contient beaucoup de paramètres à estimer ; la fonction de vraisemblance du modèle peut avoir des caractéristiques inhabituelles qui peuvent *impacter négativement l’inférence* ; déterminer le nombre de classes latentes optimales : les modèles avec des nombres de classes croissants sont testés après la sélection du premier modèle optimal basé sur l’interprétation du sujet d’étude, des approches théoriques et statistiques adéquates et des indicateurs statistiques (BIC, sample-adjusted BIC, consistent AIC, Bootstrap likelihood test, elbow plot) ; attribuer des labels à chaque cluster en utilisant les différences de leur pattern ; l’inclusion des covariables ne doit être fait qu’une fois le profil de classes optimal sélectionné, c’est-à-dire dans la solution finale directement, pour ne pas modifier le profil de solution.

Cependant, des solutions peuvent être apportées pour contrer ces limites comme par exemple le fait d’utiliser des tableaux de contingence séparés décrivant chaque transition possible pour parer la taille d’échantillon trop faible ou en assignant au préalable un ou plusieurs individus aux classes pour casser la symétrie de la distribution à postériori et atténuer les *nuisances* des densités à postériori. Ainsi, cette table contient souvent un grand nombre de patterns de réponse possibles, pouvant être vide, mais avec une taille d’échantillon plus petite que la probabilité de dispersion dans les cellules du tableau de contingence. Ou par exemple, en modifiant les fonctions de vraisemblance par l’utilisation de mesures de vraisemblance pénalisées comme l’AIC, le BIC, le Lo-Mendell-Rubin test, le bootstrapping LRT, la vérification de la distribution par prédiction Bayésienne à postériori, à la place des méthodes de test standard.

1. **GBTM (Group-Based Trajectory Modelling)**

*Description –* Le modèle GBTM est un modèle de mélange fini pour des données longitudinales continues ou catégorielles avec une distribution discrète comme une distribution de Bernouilli, de Poisson ou une distribution Normale de sa population. Ce modèle suppose qu’il n’y a pas de variation entre les individus de la même classe (pas de variance intra-class) et permet de distinguer des classes au sein d’une population homogène. Il suppose également que la variance résiduelle est constante au cours du temps et la même pour chaque classe. Pour chaque variable dépendante, il existe 2 ou plusieurs groupes avec une variation intra-groupe modéré et une variation inter-groupe élevée. Ces groupes ne peuvent pas être directement créés avec précision à partir d’une variable indépendante, c’est pourquoi une variable de classe latente catégorielle est introduit au modèle pour déterminer l’appartenance des individus au groupe. A chaque niveau de la variable latente, une trajectoire spécifique à la variable dépendante est estimée. Globalement, la proportion de la population appartenant à chaque classe est estimée, la probabilité d’appartenance aux différentes classes est calculée pour chaque individu (probabilité à postériori du groupe) puis les individus sont assignés à une classe, celle pour laquelle ils ont la plus grande probabilité d’appartenance. Pour cela, les paramètres sont estimés par maximum vraisemblance. Cette probabilité d’appartenance peut ensuite être utilisée lors de prédiction. Des covariables dépendantes ou non du temps peuvent être ajoutées au modèle.

*Objectif –* Analyser des trajectoires de groupes pour étudier leurs variations au cours du temps et pouvoir les prédire.

*Procédé –* Le modèle GBTM utilise un procédé itératif demandant des décisions à priori :

1. Hypothétiser le nombre de classes latentes attendus et la forme de la courbe pour chaque classe latente (linéaire, quadratique, cubique, …)
2. Caractérisation du modèle : dans un premier temps, il faut tester un modèle à 1 classe puis ajuster progressivement le nombre maximal de sous-groupes. Ce nombre doit être supérieur au nombre de classes attendues. Il faut pour déterminer le nombre optimal, comparer les modèles pour une trajectoire (pour 1 classe) ayant une courbe quadratique et une courbe linéaire. Puis, si le modèle quadratique est bon, comparer pour 2 trajectoires avec des composantes quadratiques. Pour tester les modèles, le critère BIC est utilisé (différences entre 2 modèles adjacents supérieures à 10) tandis que la moyenne des probabilités à postériori (supérieur à 70%) et l’entropie relative (supérieure à 0.8) sont évaluées pour tester le bon nombre de trajectoire individuelle au sein des classes identifiées
3. Estimation du modèle : il est estimé par maximum de vraisemblance ou par des méthodes Bayésiennes
4. Sélection du modèle : les connaissances liées au domaine d’étude sont une première base pour sélectionner le modèle. Il faut privilégier un modèle utile et parcimonieux pour le sujet d’étude et avoir une correspondance proche entre la probabilité de chaque classe estimée et la proportion d’individu attribués dans chaque classe selon la règle du maximum de probabilité d’appartenance. Plus précisément, la probabilité moyenne d’appartenance à postériori à une classe doit être supérieure ou égale à 0.7 et il doit y avoir au moins 5% d’individus dans chaque classe. Il doit également y avoir un intervalle de confiance raisonnable et une différence de BIC entre 2 modèles pour un nombre de trajectoire de classe différent.

*Avantages –* Le modèle GBTM est une spécification du modèle GMM plus simple car il a moins de paramètres à estimer, il est plus rapide avec moins d’erreurs ; il prend en compte les données manquantes ; accepte les résidus corrélés ; plus facile à interpréter car moins complexe et notamment plus facile à interpréter visuellement car les différences sont pertinentes entre les groupes, du point de vue clinique, et le nombre de groupes choisis est généralement faible.

*Limites –* Les données manquantes doivent être MCAR ; les groupes doivent être qualitativement différents par rapport à la variable dépendante pour qu’il n’y ait pas d’hésitation sur les différences cliniques qui existent entre les groupes ; il ne doit pas y avoir de relation directe entre les variables dépendantes et indépendantes pour que les estimations comme les odds ratios et les risques relatifs, ne soient pas systématiquement estimées ; des hypothèses fortes doivent être émises sur les distributions des trajectoires sinon le modèle pourrait aboutir à de faux résultats, fausses trajectoires et certains tests utilisés pour vérifier le modèle pourraient ne pas détecter ces erreurs ; surestimation du nombre de trajectoires quand les trajectoires individuelles ont le même profil et sont distribuées sur un continuum autour de la trajectoire moyenne ; surestimation du nombre de classes ; le modèle GBTM suppose que tous les individus d’une même classe ont le même comportement, autrement dit qu’il n’y a pas de variation intra-classe.

1. **LCA (Latent Class Analysis)**

*Description –* Le modèle LCA est un modèle de modélisation de mélanges fini, semi-paramétrique pour des données catégorielles transversales (version non longitudinale du modèle LTA). Ce modèle détermine des variables catégorielles non observées en divisant la population en classes latentes mutuellement exclusives et collectivement exhaustives. Il est surtout approprié pour une approche exploratoire. Les paramètres sont estimés par maximum vraisemblance ou par des méthodes Bayésiennes. La probabilité d’appartenance à une classe est évaluée permettant l’attribution de chaque individu à une classe. La contribution de chaque covariables peut être modélisée dans chaque classe et la probabilité d’appartenance à une classe dépend des valeurs ou des catégories des covariables. Il est préférable de recoder les variables avec le moins de catégories possibles lorsque celle-ci à un trop grand nombre de catégories à l’origine. Lorsque les données sont très nombreuses, le modèle LCA demandant des pré-requis statistiques importants, il est préférable de subdiviser des données en sous-échantillons aléatoires pour déterminer le nombre optimal de classes. Le critère BIC et la méthode elbow sont utilisés pour déterminer ce nombre optimal. Ensuite, une fois ce nombre obtenu, relancer le modèle LCA sur la population totale en fixant certains paramètres comme le nombre maximal d’itération (N = 1000) ou le nombre de fois où l’on estime le modèle avec différentes probabilités de réponses conditionnelles de classe (N = 25). Une méthode de training/testing (90%) et validation (10%) peut aussi être exécutée. Les données manquantes sont également prises en compte par des méthodes de suppression ou d’imputation multiple par exemple. Enfin, un test de χ² pour les variables catégorielles et un test de Kruskal-Wallis pour les variables continues sont appliqués pour étudier les différences entre les classes. Ces variables latentes peuvent ensuite être imputées dans des modèles de régression, par exemple, afin d’étudier les associations entre un outcome et ces populations, ses caractéristiques.

*Objectif –* Identifier des clusters non observés, non mesurés partageant des caractéristiques communes.

*Procédé –* Le modèle LCA exécute un procédé itératif basé sur des décisions à priori. Ces décisions dépendent des connaissances du chercheur dans le domaine d’étude. L’implémentation demande de nombreuses étapes :

1. Caractérisation du modèle : prendre une décision concernant les covariables additionnelles et leur codage, concernant la sélection des participants (taille d’échantillon, nombre de valeurs par participants) ;
2. Choix du nombre de classes latentes des covariables : basé sur des résultats de tests d’hypothèse ;
3. Estimation du modèle : le modèle doit être défini avant les analyses par maximum de vraisemblance (algorithme EM) ou par des méthodes Bayésiennes (algorithme MCMC). Lorsque l’échantillon est inférieur à 300 individus ou entre 300 et 500 individus, il est recommandé d’utiliser les simulations de Monte Carlo ;
4. Evaluer le nombre de classes optimal pour le modèle sélectionné : Les critères AIC, BIC, BIC ajusté sur la taille de l’échantillon, Lo-Mendell-Rubin and Vuong-Lo-Mendell-Rubin sont testés pour déterminer le meilleur modèle final. Un faible critère correspond au meilleur modèle. L’interprétation des classes est un critère important également.
5. Evaluer le modèle final : différents tests sont exécutés comme le calcul de l’entropie permettant de mesurer la séparation des classes. Une entropie proche de 1, supérieur à 0.8 correspond au meilleur modèle. Comme par exemple, le calcul de la moyenne des probabilités à postériori des classes latentes. Une fois le modèle bien défini, les covariables peuvent être ajoutées ;
6. Interpréter le modèle final : l’objectif étant de prouver sa robustesse. Des points de départ aléatoires peuvent être implémentés, des vérifications des classes ou la reproductibilité des analyses pour des autres données peuvent être analysées.

Remarque : avant de débuter les analyses, examiner les données est primordial (vérifier les valeurs extrêmes ou invraisemblables, transformer les variables : lorsqu’une variable catégorielle à moins de 10% de valeurs, exclure cette variable des analyses ou la recoder, standardiser, évaluer la corrélation des variables (>0.5, nécessite un examen particulier) pour éviter les biais et les erreurs de classifications des individus).

*Avantages –* Puissant outil pour analyser la structure des relations entre des variables catégorielles, pour explorer et interpréter des tableaux de contingences complexes, pour tester des hypothèses concernant des structures latentes de variables catégorielles ; peu d’erreurs de classification par rapport à des modèles de clustering comme le k-means ; permet de réaliser des statistiques inférentielles grâce à la visibilité des données statistiques du modèle ; plus robuste ; permet l’utilisation de données mixtes incluant des échelles différentes, pour les variables définissant les classes ; si les variables sont continues, possibilité d’utiliser des profiles et ainsi d’utiliser la méthode LPA (Latent Profile Analysis), similaire à la méthode LCA.

*Limites –* Non considérée comme une technique de modélisation de trajectoire, le modèle LTA correspondant à un modèle LCA à chaque point de temps ; coûteux donc limitation du nombre de variables selon la puissance de l’ordinateur ; sensible aux valeurs extrêmes ; pourcentage d’individus au sein des classes non connu ; beaucoup de décisions à prendre à priori.

1. **Modèle de Markov Caché**

*Description –* Le modèle de Markov caché est un modèle de statistique générative utilisé pour modéliser des processus stochastiques variant au cours du temps. Ce modèle est défini par un vecteur de probabilité initial, une matrice de transition et des densités d’émission des états cachés. Cette méthode modélise l’interaction entre 2 séquences : 1) la séquence observée correspondant à la série temporelle observée et 2) la séquence non observée correspondant à la chaine de Markov discrète, l’état caché. Autrement dit, le modèle est caractérisé par une réponse longitudinale suivant une distribution connue avec une densité qui est paramétrisée par une fonction de covariables et de coefficients de « régression » cachés dépendant de l'état. Pour la chaîne de Markov de premier ordre, chaque valeur dépend exclusivement des valeurs précédentes. À partir de l'ensemble des séquences observables, l'objectif est d’identifier le meilleur ensemble de paramètres. À partir de la séquence d'observations et de l'ensemble des paramètres du modèle de Markov caché, la séquence d’états la plus probable est estimée en utilisant l'algorithme de Viterbi. Ce modèle fait également une distinction entre un certain nombre de classes latentes qui diffèrent les unes des autres en ce qui concerne les paramètres du modèle tandis qu’à l’intérieur de chaque classe, aucune différence individuelle n’est autorisée. Le choix du nombre de classes latentes peut être défini à priori puis évalué comme étant le nombre de classe le plus raisonnable, interprétable ou alors avec des critères comme l’AIC et le BIC. La matrice de transition est, quant à elle, constituée des proportions de transition d’un état à l’autre. Les transitions inter-états sont décrites par le processus de Markov caché et la Chaîne de Markov est capable de prendre en compte différents types de données longitudinales qui sont collectées régulièrement, irrégulièrement ou de façon continue. En d’autres termes, un individu peut passer d’une classe à l’autre au cours du temps. Ces dynamismes correspondent aux trajectoires. Pour cela, une séquence de variables latentes discrètes plutôt qu’une seule variable latente est associée à chaque individu, ce qui donne lieu à un processus caché supposé suivre une chaîne de Markov. Enfin, une fois cette matrice connue, elle est appliqué à une méthode de K-Means. Après estimation du modèle par maximum de vraisemblance (algorithme EM), l’erreur standard des paramètres estimés peut être obtenu sur la base de la matrice d’information observée. Le modèle final doit être un bon compromis entre qualité et complexité. Les variables latentes doivent prendre en compte l’hétérogénéité non observée entre les individus qui ne peut pas être expliqué par les covariables observées. Une méthode de training/test et validation peut être utilisée pour déterminer le modèle final. Une fois ce modèle défini, un clustering est réalisé par la règle du maximum à postériori (MAP).

Plus spécifiquement, le modèle de Markov latent (LM) est utilisé pour des analyses de données longitudinales univariées et multivariées basées sur des observations répétées d’un échantillon au cours du temps. Ce modèle suppose que les probabilités de réponse conditionnelle sont homogènes dans le temps.

*Objectif –* Evaluer l’évolution des caractéristiques individuelles lorsque ces caractéristiques ne sont pas directement observables.

*Procédé –* Le modèle de Markov caché est un modèle à 3 états :

1. Hypothèses de Markov : la probabilité de transition d’un état à l’autre ne doit pas dépendre des états précédemment visités ou du temps écoulé au sein de l’état actuel et il est supposé que le taux de transition soit constant au cours du temps et commun à tous les individus ;
2. Estimation des paramètres du modèle : estimer les taux de transition d’une probabilité de transition qui sont directement estimables d’après les données. Une méthode Bayésienne utilisant l’algorithme MCMC peut être une solution ;
3. Inclusion des effets aléatoires et des covariables individuelles : en supposant que les taux de transition ne sont pas fixes mais modifiables ;
4. Modéliser les réponses non observées : plusieurs scénarios sont possibles en fonction des données manquantes (MAR, MCAR, MNAR, LOCF-like missingness) et des « perdus de vue ».

*Avantages –* Pour les estimations Bayésiennes, le modèle a une approche alternative très flexible et c’est une approche robuste, plus appropriée pour les petits échantillons ; l’imputation multiple Bayésienne prend en compte les données manquantes sans perte d’information ou introduction de biais (MAR) ; possibilité d’obtenir les classes latentes du model final LMM ; mise en place d’algorithmes efficaces sur le plan du calcul pour estimer et la prédire des états latents.

*Limites –* Trop peu de classes ne permet pas d’obtenir des classes occasionnelles tandis que trop de classes surestime le modèle ; la procédure d’estimation proposée ne peut pas être généralisée pour les transitions non-homogènes et l’inférence associée devient extrêmement compliqué ; le modèle ne fonctionne pas très bien lorsque le nombre d’états cachés est trop important et/ou les covariables sont de trop grande dimension ; l’estimation Bayéisenne demande de connaître à priori la distribution des paramètres du modèle ; pas de critère d'ajustement ou de sélection de modèle pour choisir le nombre de classes latentes lorsque ce nombre n'est pas connu à priori, donc l’estimation du nombre de classes latentes est calculée en utilisant l'estimation par chaîne de Markov réversible de Monte Carlo ; le changement d'étiquette impose des restrictions d'ordre sur les paramètres pour les différents états afin d'imposer un étiquetage unique ; l’état « perdu de vue » ne peut être échangé avec d’autres états.

1. **Modèle mixte**

*Description –* Le modèle mixte peut être linéaire ou plus généralisé. Dans le cas du modèle linéaire mixte, des données longitudinales ou de clusters sont analysées. Il permet d’estimer la relation entre les variables dépendantes et les effets fixes et les effets aléatoires des variables indépendantes en considérant à la fois les moyennes et les covariances. Si le nombre de prédicteurs est trop important, l’inférence et la prédiction deviennent trop complexes et infaisable. Les estimations des effets fixes peuvent être pénalisées et les effets aléatoires sélectionnés. Par exemple, les estimations des effets fixes et des effets aléatoires et la structure de la covariance des effets aléatoires, choisis simultanément dans un modèle avec une fonction de pénalisation. Cependant, pour des dimensions élevées, il est nécessaire de réduire la dimension des données avant d’utiliser le modèle car l’algorithme EM ne sera pas efficace, plausible, rapide et convergera lentement. Il est également envisageable de calculer la fonction de maximum vraisemblance et d’utiliser ensuite cet algorithme. Un second exemple est la sélection des effets fixes et aléatoires importants, indépendamment dans 2 modèles séparés. Pour cela, l’utilisation des matrices Proxy permet de prendre en compte la structure de variance-covariance inconnue des effets aléatoires durant les sélections, de minimiser l’équation de vraisemblance pénalisée pour les paramètres des effets fixes. Ensuite, la sélection des effets aléatoires est réalisée par des méthodes Bayésiennes. Pour les données à haute dimension, la dimension des effets fixes doit être plus faible que la taille d’échantillon avant utilisation du modèle. Un dernier exemple permet d’estimer et de sélectionner les effets fixes et aléatoires et la structure de covariance des effets aléatoires sélectionnés simultanément dans un modèle linéaire à effets mixtes utilisant 2 fonctions de pénalité (par une log-vraisemblance modifiée utilisant le REML (Restricted Maximum Likelihood)).

Dans le cas du modèle généralisé, c’est un modèle hybride entre le modèle linéaire mixte et le modèle linéaire généralisé. Il peut inclure des données de mesures répétées pour lesquelles l’hypothèse d’indépendance des observations est invérifiable et des variables dépendantes non Normales mesurées plus d’une fois, des données mélangées (continues et catégorielles) indépendantes grâce à la construction d’une pseudo-variable dépendante qui englobe toutes les mesures de chaque variable dépendante et la construction d’une pseudo-variable indépendante afin d'identifier l'origine de chaque instance de la pseudo-variable dépendante.

*Objectif –* Estimer la relation entre les variables dépendantes et les effets fixes et les effets aléatoires des variables indépendantes.

*Procédé –*

*Avantages –* Estimation des odds ratios et des rate ratios ; prend en compte simultanément 2, 3 variables dépendantes ou plus ; prend en compte les données manquantes.

*Limites –* Interprétation des coefficients possible si les effets aléatoires sont contrôlés par l’analyste ; même si les différences sont statistiquement significatives entre les trajectoires estimées de la variable dépendante basées sur les groupes définis par les variables indépendantes, celles-ci peuvent ne pas présenter de différences cliniquement pertinentes ; les variables non observées sont supposées MAR.

1. **Modèle joint**

*Description –* Le modèle joint modélise les processus d’observations et de mesures répétées en appliquant un modèle joint longitudinal et un modèle de Survie. Les 2 processus sont liés par le partage, spécifique à chaque individu, des effets aléatoires. Les paramètres des modèles sont estimés par un algorithme EM en traitant les effets aléatoires comme des données manquantes et ceux du modèle de Cox sont estimés par un algorithme 1-step Newton-Rasphson. Ainsi les erreurs standards peuvent être approximées après la convergence de l’algorithme EM. Concernant le modèle de Survie, conditionnellement aux effets aléatoires, le modèle pour chaque observation au cours du temps, est un modèle de risques proportionnels. Le modèle est défini en utilisant le maximum de vraisemblance. De plus, il n’est pas possible de définir une seule vraisemblance étant donné qu’elle introduit la distribution des effets aléatoires. Toutefois, des méthodes comme la quadrature Gaussienne ou l’intégration de Monte Carlo peuvent être une solution. Pour valider le modèle joint, il faut que le processus d’observations soit au moins aléatoire. Ce modèle peut facilement être généralisé à des effets aléatoires multiples, à différentes formulations paramétriques flexibles du risque de base pour le modèle des événements récurrents et pour inclure d'autres résultats. Le biais induit par le processus d’observations informatives peut être ajusté par une extension de la méthode IPW (IVW) utilisant les régressions marginales en donnant un poids pour chaque observation, basé sur l’inverse de la probabilité de chaque mesure observée. En découle une pseudo population dans laquelle le processus d’observation est statique et peut être ignoré. Les poids sont estimés par un modèle de régression incluant toutes les covariables informatives sur le processus d’observation et les dernières valeurs de chaque covariable qui pourraient impactées le processus de visite. Cependant, pour ne pas ajouter de biais, les covariables peuvent être liées au processus d'observation et doivent être incluses dans le modèle de pondération. Deux ajustements sont à réalisés car la dernière observation de chaque individu représente la fin de suivi d’étude, chaque poids est décalé d’un point de temps et étant donné que chaque individu a au moins une observation, un poids de 1 est attribué à la 1ère observation de chaque individu.

*Objectif –* Rendre compte du comportement conjoint de l’évolution d’un marqueur longitudinal quantitatif et du temps de survenue d’un événement considérant leur densité conjointe.

*Procédé –*

*Avantages –* Les modèles mixtes sont moins performants que le modèle joint ; les coefficients de régression estimés sont non biaisés car le processus de sur-modélisation des observations ne semble pas introduire de biais dans les analyses ; permet d’estimer l’association entre 2 outcomes ; des effets aléatoires additionnels peuvent être ajoutés au modèle pour expliquer l’hétérogénéité de la trajectoire des outcomes longitudinaux au cours du temps ; la forme fonctionnelle des effets du temps peut être généralisée en utilisant les *polynômes fractionnels* ou les splines ; les perdus de vues peuvent être ajoutés au sous-model de Survie ; pour la méthode QMC, accélérer le temps de calcul demande d’implanter des modèles joints multivariés avec un échantillon de petite taille et ces résultats sont comparables à l’AMC pour les échantillons de taille moyenne.

*Limites –* Pour certaines méthodes de Monte Carlo (exemple : quasi Monte Carlo), l’estimation de l’erreur MC n’est pas possible.

1. **ARIMA (Autoregression integrated moving average) & Cross Correlation**

*Description –* Le modèle ARIMA est un outil performant pour analyser des séries temporelles non-stationnaires contenant des tendances saisonnales et ordinaires, pour capturer la périodicité, la tendance et l’aléatoire des données. Il permet également l’ajout de facteurs avec des lags pour devenir alors un modèle ARIMAX (ARIMA avec des variables exogènes). Pour évaluer la corrélation entre deux séries temporelles, la cross-corrélation peut être employée. Elle l’est également pour étudier la corrélation entre 2 stimuli ou événements variant au cours du temps pouvant ou non coïncider sur des intervalles de temps. Un vecteur de mesure est sélectionné dans chaque série temporelle de telle sorte que les deux vecteurs contiennent le même nombre d'occasions puis la corrélation produit-moment de Pearson est calculée pour ces deux vecteurs. Ces vecteurs peuvent ou non démarrer en même temps. Un lag est alors introduit, c’est-à-dire un écart temporel séparant les premières mesures des 2 séries temporelles. Généralement, les données observées sont considérées comme provenant de processus stationnaires, autrement dit les moyennes et variances sont constantes au cours du temps. Pour chaque fenêtre temporelle, une windowed cross-corrélation peut être exécutée permettant d’examiner la force et les lags des associations entre les 2 séries temporelles évoluent au cours du temps. Plusieurs méthodes découlent de cette cross-corrélation, comme la DCCA qui analyse les lois des puissances de cross-corrélations entre des séries temporelles non stationnaires en supprimant les tendances locales assurant ainsi que les résultats ne sont pas affectés par ces tendances. Ou encore le modèle DPCCA qui est une approche utilisant des techniques de corrélations partielles, modèle hybride entre le DCCA et le PCCA (Partial cross-correlation analysis).

Le modèle ARIMA permet aussi de prédire l’évolution des trajectoires des séries temporelles.

*Objectif –* Analyser des séries temporelles et évaluer la corrélation entre deux séries temporelles variant au cours du temps, coïncidant ou non sur des intervalles de temps.

*Procédé –* Le modèle ARIMA et cross-corrélation est un modèle suivant dans un premier temps le processus ARIMA puis dans un second temps, il réalise une cross-corrélation :

1. Analyser le modèle ARIMA : ajuster les modèles ARIMA avec les séries temporelles, c’est-à-dire analyser la stabilité et la saisonnalité des données (ACF, PACF) puis estimer et tester le modèle (AIC et BIC, Box Ljung statistic) ;
2. Construction et validation du modèle ARIMA : séparer les données en 2 parties afin d’exécuter la méthode de training/test et validation. La première partie permettra de définir le modèle et le second de le valider ;
3. Application du modèle : validation de la prédiction et de la mise en place du modèle par RSME. La trajectoire de l’évolution des séries temporelles peut alors être analysée et prédite ;
4. Evolution des deux séries temporelles : exécution des fonctions de cross-corrélation.

*Avantages –* Une hypothèse doit être faite sur la stationnarité locale uniquement, à la place d’une hypothèse de stationnarité sur toute la durée étudiée ; des résultats robustes même lorsque des tendances non linéaires sont mélangées dans les données et des résultats qui montrent en outre des relations entre les deux données considérées sur différentes échelles de temps ; capable d'étudier les corrélations lorsque des signaux multiples sont liés ; une meilleure performance dans le traitement des corrélations dans un système complexe.

*Limites –* Les multiples signaux doivent avoir des relations linéaires entre eux mais cette limite peut être réduite par la méthode DCCA.

1. **DTW (Dynamic Time Wrapping)**

*Description –* Le modèle DTW est une approche plus flexible pour les analyses de séries temporelles incluant des données multivariées, nominales ou mixtes. Ce modèle est basé sur le concept que les similarités entre 2 séries temporelles peuvent être calculées en alignant des patterns significatifs en modifiant localement l’axe temporel dans le but de minimiser la différence cumulative entre les points alignés. Cette méthode convient aux séries temporelles ayant le même axe temporel et contenant des patterns qui sont qualitativement similaires. Cependant, les deux séries temporelles peuvent avoir des longueurs et des rythmes différents. Pour une meilleure précision du modèle et des analyses, il est préférable que les séries temporelles aient subi le moins de transformation possible et que la taille de la fenêtre temporelle ait été déterminée au préalable. La précision dépendant également de la taille de la fenêtre de temps étudiée et de la taille des données. Le principe du modèle DTW est le suivant : il regarde le meilleur alignement entre les 2 séries temporelles et estime la valeur de décalage temporelle entre les séries à chaque point de temps. Il estime également la probabilité d’observer un décalage temporelle donné à chaque point de temps lorsque le décalage actuel entre les 2 séries temporelles est de 0. Pour optimiser le modèle, il faut minimiser la distance entre les 2 séries temporelles transformées/travaillées (warpped), en utilisant par exemple le minimum des distances Euclidiennes et qui permet de plus de sélectionner des combinaisons d’alignements qui préservent l’ordre des séries temporelles.

*Objectif –* Etudier la similarité entre deux séries temporelles.

*Procédé –* Le modèle DTW est exécutable en 3 étapes :

1. Définir les dissimilarités entre les clusters : déterminées à partir de la matrice de dissimilarité individuelle par paire (ou matrice de distance) comme par exemple celles des moyennes, complètes ou de Ward ;
2. Choisir le nombre de clusters : en utilisant l’indice du coefficient de variation qui est une mesure de précision ;
3. Déterminer les centroïdes : pour cela, ce modèle utilise le calcul de la moyenne des barycentres DTW qui est une stratégie de calcul de la moyenne globale et qui fournit des centroïdes pour des groupes de séries temporelles bien adaptés à la dissimilarité DTW. Ce calcul dépend de l’étape initiale, des données initiales.

*Avantages –* Le modèle DTW accepte les séries temporelles incomplètes avec des références complètes, grâce à la méthode d’open-end DTW ; accepte les trajectoires avec des nombres de points différents ; rapide à exécuter et efficace pour la calcul du décalage temporel entre 2 séries temporelles, même lorsque la proportion de bruit correspond à 20-30% de la variance totale ; performant pour les profils comprenant des modifications récurrentes et non récurrentes et permet d’estimer la variation en amplitude et direction du décalage temporel.

*Limites –* Le modèle DTW a des contraintes au niveau des données de début et de fin d’étude, c’est-à-dire que les deux séries temporelles doivent entièrement couvrir les mêmes zones temporelles, et des contraintes au niveau de la pente locale de la trajectoire des séries temporelles ; il ne doit pas y avoir de boucle temporelle (monoticité) ; coût de calcul élevé ; a besoin d’un processus spécifique de calcul des moyennes pour créer les centroïdes des clusters et exige que les séries temporelles d'essai et de référence soient connues dans leur intégralité afin de fournir une distance.

D’autres méthodes existent comme les méthodes de clustering (cross-sectional data, continuous, categorical, mixed data), les méthodes d’analyse de séquence (categorical longitudinal data to describe the sequence of events), … (Lore et al., 2020), K-means for longitudinal data, change indices + PCA + k-means (Verboon, 2022), Factor Mixture Analysis (FMA), k-means, Mixture Regression Analysis, Configural Frequency Analysis (Hofmans, 2020), latent growth curve modeling (Charnigo, 2011) => Mais pas forcément idéales pour les données longitudinales/télémonitoring/de trajectoire

# II/ Comment s’en servir ? Quel choix faire ?

Selon les caractéristiques des données (variables catégorielles ou continues, le nombre de mesures temporelles, la taille d’échantillon, le but de l’étude, le type de données manquantes, etc.), certains modèles statistiques sont plus appropriés que d’autres. Ce choix est important pour l’obtention de résultats mais il est difficile au vu du nombre important de méthodes développées. Nous proposons dans cette partie, d’associer les méthodes listées auparavant aux types de données rencontrées. Selon plusieurs critères, nous devons opter pour l’une ou l’autre méthode : données manquantes, durée des données, perdus de vues, standardisation des variables, …

*Le modèle LTA a quelques similarités avec le modèle de Markov caché.*

Tableau de synthèse

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Méthode** | **Objectif** | **Avantages** | **Limites** | **Pour quel type de données ?** | **Pour quelles études ?** | **Quand l’utiliser ?** | **Package R** |
| GMM (Growth Mixture Modelling) | Examiner des trajectoires uniques de données longitudinales avec des mesures de suivi répétées ; identifier des patterns de trajectoire ; identifier des classes non observées et décrire les changements longitudinaux au sein de chacune de ces classes non observées | Prend en compte les données manquantes ; accepte les résidus corrélés ; capture les différences inter et intra individus au cours du temps ; c’est un modèle approprié pour modéliser la stabilité et les changements au cours du temps ; permet à la trajectoire d’être qualitativement différente au cours du temps selon les différents groupes d’individus | Estime beaucoup de paramètres ; son interprétation peut être complexe ; certains paramètres doivent être choisis à priori ; c’est une technique exploratoire contrainte qui peut identifier de faux groupes | Continues ou catégorielles | Etudes longitudinales avec des mesures de suivi répétées dans le temps |  | *gmm* |
| LTA (Latent Transition Analysis) | Etudier la probabilité de transition d’une classe à un point de temps, à une autre classe au point de temps suivant | Modélise un changement au cours du temps et identifie les prédicteurs de ces changements ; compare les différentes classes pour déterminer leurs caractéristiques et pour évaluer la contribution des différentes mesures pour chaque classe latente ; il est approprié pour modéliser la stabilité et le changement temporel par sa capacité à évaluer les différences inter-individu dans les processus intra-individu | Large échantillon d’individu car le modèle doit estimer de nombreux paramètres et utilise généralement le ‘burn-in’; complexité du modèle en augmentant le nombre de points de temps (> 6); la fonction de vraisemblance du modèle peut avoir des caractéristiques inhabituelles qui peuvent *impacter négativement l’inférence* ; déterminer le nombre de classes latentes optimales ; attribution des labels à chaque cluster ; l’inclusion des covariables | Catégorielles, tableaux de contingence | Etudes longitudinales | Le modèle LTA correspondant à un modèle LCA à chaque point de temps | *LMest* |
| GBTM (Group-Based Trajectory Modelling) | Analyser des trajectoires de clusters pour étudier leurs variations au cours du temps et pouvoir les prédire | Plus simple que le modèle GMM car il a moins de paramètres à estimer, il est plus rapide avec moins d’erreurs ; il prend en compte les données manquantes ; accepte les résidus corrélés ; plus facile à interpréter, notamment visuellement, car moins complexe | Les données manquantes doivent être MCAR ; les groupes doivent être qualitativement différents par rapport à la variable dépendante; il ne doit pas y avoir de relation directe entre les variables dépendantes et indépendantes ; des hypothèses fortes doivent être émises sur les distributions des trajectoires ; surestimation du nombre de trajectoires quand les trajectoires individuelles ont le même profil et sont distribuées sur un continuum autour de la trajectoire moyenne ; surestimation du nombre de classes ; le modèle GBTM suppose que tous les individus d’une même classe ont le même comportement (pas de variation intra-classe) | Continues ou catégorielles, avec une densité discrète | Etudes longitudinales | Lorsque le modèle GMM ne fonctionne pas, devient trop complexe à exécuter | *gbmt* |
| LCA (Latent Class Analysis) | Identifier des clusters non observés, non mesurés partageant des caractéristiques communes | Puissant outil pour analyser la structure des relations entre des variables catégorielles, pour explorer et interpréter des tableaux de contingences complexes, pour tester des hypothèses concernant des structures latentes de variables catégorielles ; peu d’erreurs de classification; permet de réaliser des statistiques inférentielles; plus robuste ; permet l’utilisation de données mixtes incluant des échelles différentes, pour les variables définissant les classes ; si les variables sont continues, possibilité d’utiliser des profiles et ainsi d’utiliser la méthode LPA | Coûteux donc limitation du nombre de variables selon la puissance de l’ordinateur ; sensible aux valeurs extrêmes ; pourcentage d’individus au sein des classes non connu ; beaucoup de décisions à prendre à priori | Catégorielles | Etudes transversales | Ne pas utiliser lorsque l’objectif est de modéliser une trajectoire | *poLCA* |
| Markov Caché | Evaluer l’évolution des caractéristiques individuelles lorsque ces caractéristiques ne sont pas directement observables | Pour les estimations Bayésiennes, le modèle a une approche alternative très flexible et c’est une approche robuste, plus appropriée pour les petits échantillons ; l’imputation multiple Bayésienne prend en compte les données manquantes sans perte d’information ou introduction de biais (MAR) ; possibilité d’obtenir les classes latentes du model final LMM ; algorithmes efficaces | Trop peu de classes ne permet pas d’obtenir des classes occasionnelles tandis que trop de classes surestime le modèle ; la procédure d’estimation proposée ne peut pas être généralisée pour les transitions non-homogènes et l’inférence associée devient extrêmement compliqué ; le modèle ne fonctionne pas très bien lorsque le nombre d’états cachés est trop important et/ou les covariables sont de trop grande dimension ; l’estimation Bayéisenne demande de connaître à priori la distribution des paramètres du modèle ; pas de critère d'ajustement ou de sélection de modèle pour choisir le nombre de classes latentes, l’estimation du nombre de classes latentes est calculée en utilisant l'estimation par chaîne de Markov réversible de Monte Carlo ; le changement d'étiquette impose des restrictions d'ordre sur les paramètres pour les différents états; l’état « perdu de vue » ne peut être échangé avec d’autres états | Univariées, multivariées | Etudes longitudinales, processus stochastiques variant au cours du temps, 1 série temporelle observée et 1 série temporelle cachée (?) |  | *Voir Midelet article ?* |
| Mixte | Estimer la relation entre les variables dépendantes et les effets fixes et les effets aléatoires des variables indépendantes | Estimation des odds ratios et des rate ratios ; prend en compte simultanément 2, 3 variables dépendantes ou plus ; prend en compte les données manquantes | Interprétation des coefficients possible si les effets aléatoires sont contrôlés par l’analyste ; même si les différences sont statistiquement significatives entre les trajectoires estimées de la variable dépendante, celles-ci peuvent ne pas présenter de différences cliniquement pertinentes ; les variables non observées sont supposées MAR | Mesures répétées dans le temps | Etudes longitudinales ou de clusters |  | *lme4* |
| Joint | Rendre compte du comportement conjoint de l’évolution d’un marqueur longitudinal quantitatif et du temps de survenue d’un événement considérant leur densité conjointe | Les coefficients de régression estimés sont non biaisés ; permet d’estimer l’association entre 2 outcomes ; des effets aléatoires additionnels peuvent être ajoutés au modèle ; la forme fonctionnelle des effets du temps peut être généralisée en utilisant les *polynômes fractionnels* ou les splines ; les perdus de vues peuvent être ajoutés au sous-model de Survie | Pour certaines méthodes de Monte Carlo (exemple : quasi Monte Carlo), l’estimation de l’erreur MC n’est pas possible | Continues, catégorielles | Processus d’observations et de mesures répétées, études longitudinales | Lorsque les décès du modèle de survie ne sont pas assez nombreux, privilégier le modèle mixte mais dans les autres cas privilégier le modèle joint car plus performant | *JM* |
| ARIMA (AutoRegression Integrated Moving Average) & Cross-correlation | Analyser des séries temporelles et évaluer la corrélation entre deux séries temporelles variant au cours du temps, coïncidant ou non sur des intervalles de temps | Une hypothèse doit être faite sur la stationnarité locale uniquement ; des résultats robustes même lorsque des tendances non linéaires sont mélangées dans les données et des résultats qui montrent en outre des relations entre les deux données considérées sur différentes échelles de temps ; capable d'étudier les corrélations lorsque des signaux multiples sont liés ; une meilleure performance dans le traitement des corrélations dans un système complexe | Les multiples signaux doivent avoir des relations linéaires entre eux | Quantitatives discrètes ou continues | Séries temporelles non-stationnaires contenant des tendances saisonnières et ordinaires |  | *forecast, tseries,* |
| DTW (Dynamic Time Warping) | Regarder la similarité entre deux séries temporelles | Accepte les séries temporelles incomplètes avec des références complètes ; accepte les trajectoires avec des nombres de points différents ; rapide à exécuter et efficace pour le calcul du décalage temporel entre 2 séries temporelles ; performant pour les profils comprenant des modifications récurrentes et non récurrentes et permet d’estimer la variation en amplitude et direction du décalage temporel | Les deux séries temporelles doivent entièrement couvrir les mêmes zones temporelles ; contraintes au niveau de la pente locale de la trajectoire des séries temporelles ; il ne doit pas y avoir de boucle temporelle (monoticité) ; coût de calcul élevé ; a besoin d’un processus spécifique de calcul des moyennes pour créer les centroïdes des clusters et exige que les séries temporelles d'essai et de référence soient connues dans leur intégralité | Multivariées, nominales ou mixtes | Séries temporelles |  | *Dtw, TimeShift* |

# III/ Exemple sur nos données (?)

Expliquer la base de données, quelle(s) méthode(s) on a choisi et pourquoi + quelques résultats (?)

1. E-Meuse
2. Agir A Dom’, MARS

# Recherche PUBMED/InsermBiblio

Donner les mots-clefs utilisés + faire un tableau récapitulatif de la biblio (titre, auteur, date, méthode utilisée, type de données, objectif)

# Références

1. Liu X. Introduction [Internet]. In: Methods and Applications of Longitudinal Data Analysis. Elsevier; 2016 [cited 2024 Dec 20]. p. 1–18.Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128013427000010

2. Dusart C, Andre S, Mettay T, Bruyneel M. Telemonitoring for the Follow-Up of Obstructive Sleep Apnea Patients Treated with CPAP: Accuracy and Impact on Therapy. *Sensors (Basel)* 2022;22(7):2782.

1. La réponse à l’item est le pattern de réponse des données observés et de la variable latente pour chaque classe latente. [↑](#footnote-ref-1)