Optimisation de méthodes de surveillance épidémiologique multi-sources : Approches statistiques innovantes et intégration dynamique pour la production d'indicateurs en temps réel.

Mots-clés :
- Surveillance épidémiologique
- Intégration de données hétérogènes
- Modélisation spatio-temporelle
- Régularisation adaptative
- Modèles prédictifs

Unité/équipe encadrante :
LTSI INSERM U1099, Équipe DOMASIA
2 rue du Professeur Léon Bernard

Directeurs de thèse :

Morgane Pierre-Jean (Ingénieure de Recherche- Data Scientist, PhD)
Valérie Bertaud (PU-PH)

Contact:
Morgane.pierre-jean@univ-rennes.fr

Résumé :
La surveillance épidémiologique est un pilier essentiel de la santé publique, permettant de détecter précocement les épidémies et de mettre en place des mesures de contrôle efficaces. Les systèmes traditionnels, basés principalement sur des réseaux sentinelles, présentent cependant des limitations en termes de réactivité, avec des délais pouvant atteindre plusieurs semaines avant que les données ne soient disponibles et analysées.
Cette thèse vise à développer un cadre méthodologique innovant pour optimiser les méthodes de surveillance épidémiologique en exploitant de multiples sources de données hétérogènes. En intégrant des données issues des entrepôts de données de santé (EDS), des informations provenant du web (recherches sur les moteurs de recherche, réseaux sociaux), des données environnementales (conditions météorologiques, pollution atmosphérique) et d'autres sources pertinentes, nous cherchons à améliorer la précision et la rapidité des indicateurs prédictifs.
Notre approche s'appuiera sur l'optimisation et l'extension de modèles statistiques avancés, tels que les modèles autorégressifs pénalisés (ex. ARGO, ARGONet), en développant de nouvelles méthodes de régularisation adaptative pour la sélection dynamique de variables. Nous explorerons également l'intégration explicite des dépendances spatio-temporelles pour mieux modéliser la propagation géographique des épidémies. Les méthodes développées seront validées sur différentes pathologies, notamment les infections respiratoires aiguës (IRA), les infections sexuellement transmissibles (IST) et la dengue, dans le cadre du projet **ORCHIDEE (**Organisation d'un

Réseau de Centres Hospitaliers Impliqués Dans la surveillance Epidémiologique et la réponse aux Emergences).

Contexte scientifique :

La surveillance épidémiologique moderne nécessite une adaptation aux défis posés par l'augmentation des flux de données et la nécessité d'une réaction rapide face aux épidémies émergentes. L'intégration de sources de données multiples et hétérogènes offre un potentiel considérable pour améliorer la détection précoce et la prévision des épidémies. Cependant, plusieurs défis méthodologiques doivent être relevés :

1. **Sélection dynamique et adaptative de variables :**

Les méthodes traditionnelles de sélection de variables ne tiennent pas compte de la nature dynamique des données épidémiologiques. Il est nécessaire de développer des techniques de régularisation adaptative qui s'ajustent aux variations temporelles et saisonnières, permettant une sélection de variables qui évolue avec les tendances des données.

2. **Fusion efficace de sources hétérogènes :**

Les différentes sources de données présentent des caractéristiques variées en termes de fréquence, de qualité, de retard et de pertinence. Développer des stratégies de fusion dynamique qui calibrent de manière optimale les contributions de chaque source est crucial pour obtenir des prédictions robustes et précises.

3. **Modélisation spatio-temporelle avancée :**

Les épidémies se propagent non seulement dans le temps mais aussi dans l'espace. L'intégration des relations spatiales et des corrélations interrégionales dans les modèles statistiques est essentielle pour comprendre et prévoir la diffusion géographique des maladies.

Questions scientifiques :

1. **Optimisation de la sélection de variables :**
    - Quelle approche de régularisation adaptative permettrait d'intégrer efficacement les contraintes temporelles et de s'adapter aux variations saisonnières ?
    - Comment généraliser ces méthodes pour les appliquer à diverses pathologies aux dynamiques épidémiologiques différentes ?
2. **Architecture de fusion dynamique :**
    - Comment calibrer de manière optimale les contributions de multiples sources de données hétérogènes en temps réel ?
    - Quelle stratégie de pondération adaptative peut être mise en place pour refléter les variations de qualité et de pertinence des différentes sources au cours du temps ?
    - Comment intégrer les incertitudes associées à chaque source de données dans le processus de fusion pour améliorer la robustesse des prédictions ?
3. **Intégration spatio-temporelle :**
    - Comment étendre les modèles existants pour capturer les dépendances spatiales et temporelles dans la propagation des épidémies ?

- Quelle approche de régularisation spatio-temporelle peut être développée pour modéliser efficacement les corrélations interrégionales et la propagation géographique des maladies ?

Grandes étapes :

1. **Analyse théorique et état de l'art (6 mois) :**
   - Effectuer une revue exhaustive des approches existantes en matière de sélection de variables, de fusion de données hétérogènes et de modélisation spatio-temporelle dans le contexte de la surveillance épidémiologique.
   - Identifier les limitations des méthodes actuelles et déterminer les axes d'amélioration potentiels.
   - Définir un cadre d'évaluation pour comparer les nouvelles méthodes développées aux approches existantes.
2. **Développements méthodologiques (18 mois) :**
   - **Régularisation adaptative :** Développer des méthodes de sélection dynamique de variables, capables de s'ajuster aux variations temporelles et saisonnières des données, en s'appuyant sur des techniques telles que le LASSO adaptatif ou la régularisation par pénalités variables.
   - **Fusion dynamique multi-sources :** Concevoir des algorithmes qui calibrent en temps réel les contributions des différentes sources de données, en intégrant des techniques d'apprentissage automatique et de théorie de l'information.
   - **Modélisation spatio-temporelle :** Élaborer des modèles intégrant explicitement les relations spatiales et les corrélations interrégionales, en utilisant des approches telles que les modèles à effets mixtes spatio-temporels ou les processus gaussiens.
3. **Validation et application (12 mois) :**
   - **Validation sur données synthétiques :** Tester et valider les développements méthodologiques sur des jeux de données simulées pour évaluer leur performance et leur robustesse.
   - **Application aux données réelles :** Appliquer les méthodes développées à des données réelles issues de différentes pathologies, telles que les infections respiratoires aiguës (IRA), les infections sexuellement transmissibles (IST) et la dengue, en collaboration avec les partenaires du projet ORCHIDEE.
   - **Comparaison avec les approches existantes :** Évaluer les gains en termes de précision, de réactivité et de robustesse des prédictions par rapport aux méthodes traditionnelles.
4. **Dissémination des résultats et rédaction de la thèse (6 mois) :**
   - **Publications scientifiques :** Rédiger des articles pour des revues internationales à comité de lecture et présenter les résultats lors de conférences spécialisées.
   - **Rédaction de la thèse :** Synthétiser les contributions scientifiques et les résultats obtenus pour la rédaction du manuscrit de thèse.

Impact attendu :
Les méthodes développées contribueront à améliorer significativement la surveillance épidémiologique en fournissant des outils prédictifs plus précis et réactifs. En intégrant efficacement des sources de données hétérogènes et en tenant compte des dynamiques spatio-temporelles, ces approches permettront aux autorités sanitaires de détecter plus rapidement les épidémies émergentes et de mettre en place des mesures de contrôle adaptées. Cela aura un impact direct sur la gestion des crises sanitaires et la protection de la population.

Compétences requises :

- **Statistiques et modélisation :** Solides connaissances en statistiques, en particulier en modélisation statistique et en méthodes de sélection de variables.
- **Apprentissage automatique :** Compétences en machine learning, y compris l'expérience avec les modèles prédictifs et les algorithmes d'apprentissage.
- **Traitement de données massives :** Expérience dans la manipulation et l'analyse de grands ensembles de données hétérogènes.
- **Programmation :** Maîtrise des langages de programmation tels que R ou Python, et des bibliothèques associées (scikit-learn, TensorFlow, etc.).
- **Connaissances en épidémiologie :** Compréhension des principes de base de l'épidémiologie et de la santé publique.

Références bibliographiques :

Sylvestre E, Cécilia-Joseph E, Bouzillé G, Najioullah F, Etienne M, Malouines F, Rosine J, Julié S, Cabié A, Cuggia M. The Role of Heterogenous Real-world Data for Dengue Surveillance in Martinique: Observational Retrospective Study. JMIR Public Health Surveill. 2022 Dec 22;8(12):e37122. doi: 10.2196/37122. PMID: 36548023; PMCID: PMC9816958.

Liu D, Clemente L, Poirier C, Ding X, Chinazzi M, Davis J, Vespignani A, Santillana M
Real-Time Forecasting of the COVID-19 Outbreak in Chinese Provinces: Machine Learning Approach Using Novel Digital Data and Estimates From Mechanistic Models
J Med Internet Res 2020;22(8):e20285
URL: https://www.jmir.org/2020/8/e20285
DOI: 10.2196/20285

Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert ML, Chabot M, Chazard E, Lavenu A, Cuggia M. Leveraging hospital big data to monitor flu epidemics. Comput Methods Programs Biomed. 2018 Feb;154:153-160. doi: 10.1016/j.cmpb.2017.11.012. Epub 2017 Nov 15. PMID: 29249339.

Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, Bouzillé G. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study. JMIR Public Health Surveill. 2018 Dec 21;4(4):e11361. doi: 10.2196/11361. PMID: 30578212; PMCID: PMC6320394.