

Intitulé du stage : Stage de Data engineering/Science pour la surveillance en santé publique (SurSaUD)

Référence de l'offre : DATA-STA-2026-08

■ Présentation de Santé publique France

Santé publique France est l'agence nationale de santé publique française. Etablissement public de l'Etat sous tutelle du ministre chargé de la santé créé par l'ordonnance 2016-246 du 15 avril 2016, elle intervient au service de la santé des populations. Agence scientifique et d'expertise du champ sanitaire, elle a pour missions :

- 1° L'observation épidémiologique et la surveillance de l'état de santé des populations ;
- 2° La veille sur les risques sanitaires menaçant les populations ;
- 3° La promotion de la santé et la réduction des risques pour la santé ;
- 4° Le développement de la prévention et de l'éducation pour la santé ;
- 5° La préparation et la réponse aux menaces, alertes et crises sanitaires ;
- 6° Le lancement de l'alerte sanitaire.

L'agence est organisée autour de 12 directions scientifiques, transversales ou assurant le soutien à l'activité.

Les orientations stratégiques de l'agence et son programme de travail, arrêtés par son Conseil d'administration, se déclinent en trois axes : Consolider la capacité d'anticipation et de réponse réactive pour faire face aux menaces sanitaires ; Mesurer et évaluer l'ampleur des maladies et des facteurs de risques pour guider leur prévention et leur contrôle ; Renforcer l'impact sur la santé dans toutes les politiques publiques et la prévention et promotion de la santé.

Le siège de l'agence est situé à Saint-Maurice (94). L'agence dispose d'implantations régionales auprès des agences régionales de la santé.

Localisation du stage :

- ☒ Saint-Maurice
☐ En région, indiquer la région et la ville concernées :

■ Le stage

Contexte :

Description de la Direction : Le stage s'inscrit au sein de la Direction Appui, Traitements et Analyses des données (DATA) de Santé publique France, qui assure un appui transversal à l'ensemble de l'agence pour le traitement, l'analyse et la valorisation des données. Forte d'une cinquantaine d'agents, la direction est structurée en trois unités spécialisées (unité « Applications, big data et surveillance syndromique », unité « Appui et méthodes pour les études et investigations dans le domaine de la surveillance », unité « Enquêtes »). Elle intervient sur l'ensemble du cycle de vie des données de santé.

Ses missions couvrent notamment la gestion de données, l'analyse statistique, la géomatique, la métrologie, ainsi que le développement d'outils informatiques d'analyse et de visualisation. Elle pilote ou soutient plusieurs dispositifs structurants, notamment le système de surveillance syndromique SurSaUD, l'enquête Baromètre santé, le site open-DATA Odyssée ainsi que l'exploitation de bases médico-administratives comme le SNDS.

La direction DATA développe une expertise reconnue en modélisation spatio-temporelle, détection automatique de signaux, et intelligence artificielle. Soucieuse de renforcer la qualité scientifique de ses travaux, elle collabore activement avec des partenaires institutionnels et académiques, et accueille régulièrement stagiaires, interne, doctorants et chercheurs, dans une dynamique d'innovation continue au service de la surveillance en santé publique.

Description de l'Unité : L'unité AMETIS (Appui et Méthodes pour les Études et Investigations dans le domaine de la Surveillance) met en œuvre des méthodes quantitatives avancées et des outils d'ingénierie / science des données pour accompagner les études et dispositifs de surveillance de Santé publique France. Elle couvre l'intégralité du cycle des données, avec une expertise structurée autour de cinq axes principaux :

- Conception d'études : élaboration des designs méthodologiques et des protocoles de recueil de données.
- Ingénierie des données : nettoyage, structuration et automatisation de pipelines, ainsi que gestion de bases de données complexes.
- Modélisation et science des données : mise en œuvre d'analyses statistiques avancées et de méthodes de data science pour produire des indicateurs épidémiologiques fiables et innovants
- Visualisation et outils interactifs : Développement de cartographies dynamiques, de tableaux de bord interactifs et de rapports automatisés.
- Veille et innovation : Intégration de méthodes émergentes, IA et collaborations académiques.

En parallèle, AMETIS contribue aux dispositifs nationaux (ex : portail open-DATA Odissé) en assurant la qualité, la traçabilité et la reproductibilité des données et analyses produites

Sujet du stage : Ce stage a pour objectif de contribuer à la mise en place d'une chaîne automatisée, à grande échelle, de traitement et de structuration des données issues du système SurSaUD, adaptée à l'analyse de séries temporelles multivariées en contexte de surveillance sanitaire (détection de signaux faibles, modélisation, prévision). Le ou la stagiaire participera à la fiabilisation, l'enrichissement et l'organisation d'un jeu de données massif, constitué de plusieurs centaines de milliers de séries temporelles, décrivant l'évolution d'indicateurs de santé selon différentes dimensions géographiques, temporelles, syndromiques et démographiques.

Missions du stage : Les tâches spécifiques du stagiaire seront précisées selon le profil du candidat. Le stage portera notamment sur les missions suivantes :

Pour la structuration et l'automatisation des données, le-la stagiaire sera amené-e à participer aux tâches suivantes :

- **Collecte et structuration des données** : mobiliser les différentes sources du système SurSaUD et leurs référentiels associés pour constituer un jeu de données cohérent et structuré.
- **Constitution d'un jeu de données de référence** : participer à l'élaboration d'un corpus structuré et documenté, destiné à l'évaluation comparative de méthodes de détection de signaux et de prévision.
- **Construction d'un référentiel de séries temporelles** : organiser les données en un corpus structuré selon plusieurs dimensions (géographiques, temporelles, syndromiques et démographiques), adapté aux analyses statistiques et exploratoires.
- **Enrichissement contextuel** : intégrer des données externes (caractéristiques sociodémographiques, calendrier scolaire, jours fériés, conditions météorologiques...) pour affiner l'analyse des séries et orienter les choix méthodologiques futurs (pré-sélection de variables explicatives).
- **Contrôle qualité** : mettre en place des procédures d'évaluation de la qualité des données incluant la détection et, si nécessaire, l'imputation des valeurs manquantes, l'identification des anomalies, ainsi que la définition d'indicateurs synthétiques de fiabilité.
- **Documentation des traitements** : produire une documentation technique complète (guides, scripts, structuration des jeux de données) et mettre en œuvre des pratiques garantissant la reproductibilité et la généralisation des travaux à d'autres sources de données (versioning, modularité des scripts).
- **Optimisation des performances** : adapter les traitements aux infrastructures de calcul de Santé publique France, en tenant compte des contraintes liées aux volumes importants de données manipulées.
- **Automatisation des traitements** : concevoir et implémenter une chaîne de traitement automatisée, garantissant la traçabilité, l'horodatage et la reproductibilité des processus.

Pour l'analyse et la valorisation des données, le-la stagiaire sera amené-e à participer aux tâches suivantes :

- **Extraction de descripteurs temporels** : compiler des indicateurs décrivant le comportement des séries (statistiques globales, saisonnalité, tendance, variabilité...) en vue d'une typologie des dynamiques observées.

- **Analyse des corrélations et dépendances temporelles** : étudier les relations entre les séries temporelles pour identifier des interactions ou des dépendances significatives entre indicateurs de santé ou dimensions géographiques.
- **Analyse exploratoire et classification non supervisée** : regrouper les séries selon leurs profils temporels afin d'identifier des tendances récurrentes ou atypiques et de mieux caractériser la diversité des comportements.

Environnement de travail : Le stage s'intègre dans un environnement technique dynamique et collaboratif, mobilisant des outils de développement modernes, des langages adaptés à la science des données, et des infrastructures de calcul performantes. Le-la stagiaire évoluera au sein d'une équipe pluridisciplinaire, en interaction étroite avec des épidémiologistes, data scientists, statisticiens, ingénieurs et membres de la DSI ainsi que le RSSI.

Les principaux outils et technologies mobilisés incluent :

- Langages : Python, R
- Environnement collaboratif : GitLab (versionning, intégration continue, gestion des issues)
- Automatisation et orchestration : Apache Airflow (déploiement, supervision des workflows), Docker
- Formats et bases de données : PostgreSQL, DuckDB, fichiers Parquet, CSV
- Environnements de développement : VS Code, Positron, IA Mistral
- Traitement intensif : Apache Spark, via les serveurs de calcul internes de Santé publique France

Profil recherché : Nous recherchons un-e stagiaire de niveau Master 2, idéalement en informatique, data science, statistique, mathématiques appliquées ou disciplines connexes. Le-la candidat-e doit avoir une bonne connaissance du langage Python et de R et avoir des connaissances en développement logiciel, notamment sur les bonnes pratiques telles que les tests unitaires, le versionning et l'intégration continue. Une expérience avec des outils comme GitLab, Airflow ou PostgreSQL serait un plus.

Le-la candidat-e devra faire preuve de bonnes capacités d'analyse, d'adaptation et d'organisation, ainsi que d'un réel esprit d'équipe pour évoluer au sein d'un environnement pluridisciplinaire. Ce stage offre l'opportunité de participer à des projets techniques innovants, au service de la surveillance de la santé publique, en étroite collaboration avec des experts en épidémiologie, data science et ingénierie, dans un cadre de travail stimulant, bienveillant et collaboratif.

Description :

Pour tout renseignement complémentaire, consulter le portail sur le site <http://www.santepubliquefrance.fr>

▪ Type de stage proposé

- | | | | |
|--|--|---|---|
| ☐ Licence | ☐ Master 1 | ☐ Master 1 + sujet tutoré* | ☐ Master 1 « observation »** |
| ☒ Master 2 Professionnel | ☒ Master 2 Recherche | | |

* stage M1 + sujet tutoré (proposé par l'ISPED) : le stage complète un travail préalable de sujet tutoré de 2 mois et demi (revue de la littérature, analyse d'une base de données, etc.)

** stage M1 « observation » (proposé par l'ISPED) : le stagiaire contribue aux activités courantes de l'équipe sans être en charge d'un dossier particulier

Gratification du stage : selon la réglementation en vigueur dans les établissements publics

▪ Date proposée pour le stage et durée :

- ☒ Sans contrainte de date
- ☐ A partir de : (indiquer une date)

Durée minimum : 6 mois

Extension possible au-delà de la période obligatoire ☒ Oui ☐ Non

▪ Prérequis

- ☐ Aucun
- ☒ Compétences spécifiques : Connaissance de l'ingénierie de données et des langages Python / R

☒ Maîtrise d'un logiciel spécifique : Langage Python et/ou R

☒ Autre (préciser) : Connaissance sur les outils de versionning de code tels que Git/GitHub/GitLab

▪ **Stage proposé par**

Direction : DATA

Unité : AMETIS

Maître(s) de stage / personne contact :

Nom : HANF

Prénom : MATTHIEU

Fonction : Directeur DATA

Téléphone : 01 71 80 16 86

Adresse email : matthieu.hanf@santepubliquefrance.fr

Co-encadrement

Nom : FOUILLET

Prénom : Anne

Fonction : Coordinatrice SURSAUD

Téléphone :

Adresse email : anne.fouillet@santepubliquefrance.fr

Co-encadrement

Nom : SOMMEN

Prénom : Cécile

Fonction : Data Scientist Senior

Téléphone : 01 41 79 57 38

Adresse email : camille.pelat@santepubliquefrance.fr

▪ **Pour postuler**

Adresser les candidatures (lettre de motivation + cv) en indiquant la référence de l'annonce par courriel : recrut@santepubliquefrance.fr