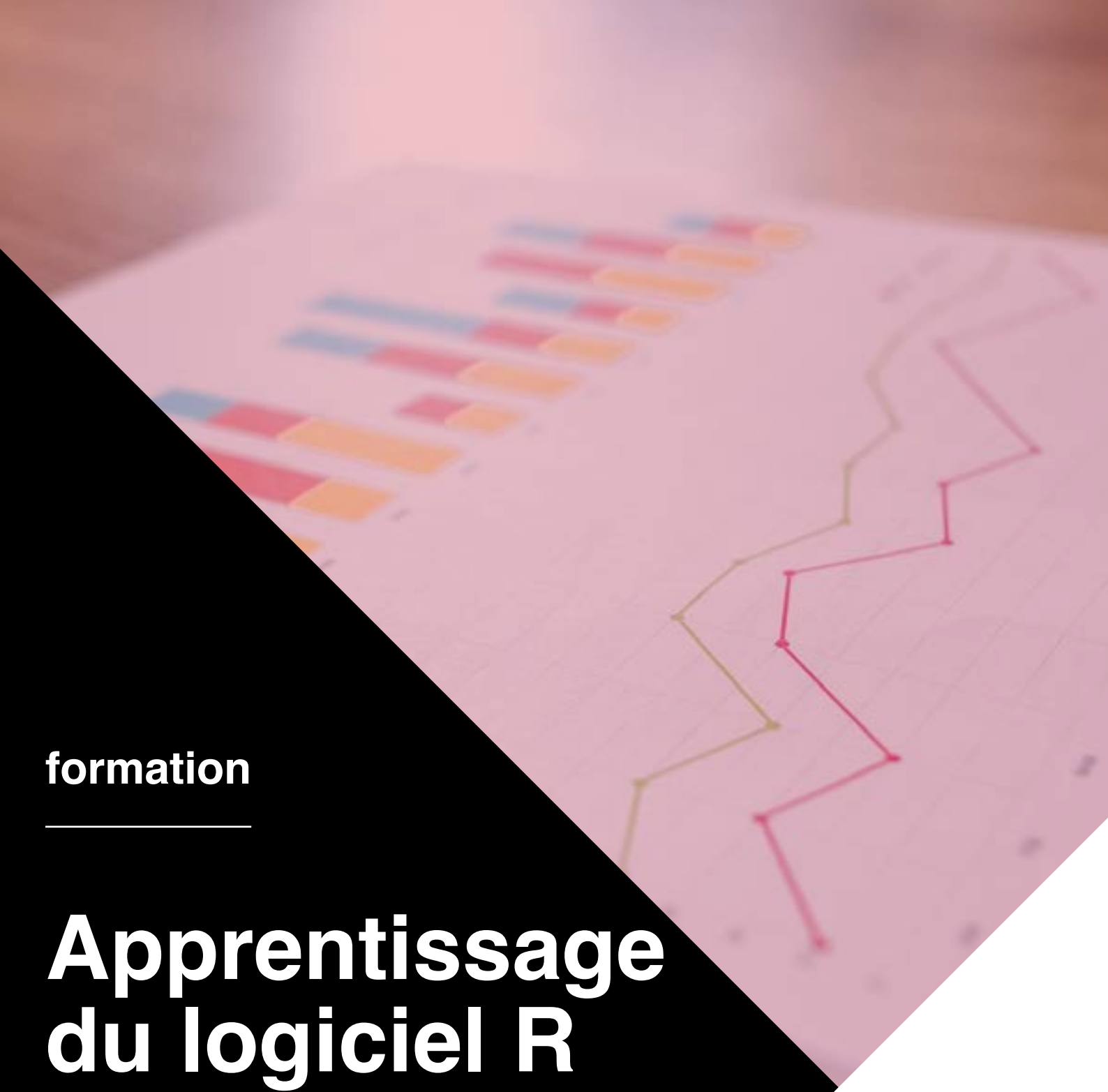

formation

# Apprentissage du logiciel R

2023-2024

www.ephe.psl.eu

École Pratique des Hautes Études | PSL

# Formation « Apprentissage du logiciel R »

## Présentation

Le logiciel libre R est un logiciel **extrêmement utilisé pour l'analyse statistique des données.** Cependant, avant tout type d'analyses statistiques, il est important d'importer correctement les données, en visualiser le contenu et les transformer dans un format adéquat. Ces étapes, **souvent peu abordées** dans les formations, sont pourtant **cruciales** à deux titres. Premièrement, en fonction de la complexité du jeu de données originel, elles peuvent **constituer l'essentiel du temps** total consacré au traitement des données. La connaissance des bons outils peut **transformer un travail laborieux de plusieurs heures en une tâche de quelques dizaines de minutes**. Deuxièmement, l'utilisation d'outils « clique-bouton » (type tableur Excel) lors de ces étapes entraîne un fort risque d'introduction d'erreurs humaines et d'une difficulté à reproduire à nouveau le même travail simplement. Les outils que propose le langage R permettent **d'optimiser, d'automatiser, de rendre reproductible et facilement communicable** toutes ces étapes cruciales, mais ils sont trop peu souvent enseignés, au profit d'un contenu plus statistique.

Cette formation propose donc de se focaliser sur toutes ces étapes préliminaires et complémentaires à une analyse statistique pour permettre aux apprenants **d'abandonner le travail laborieux effectué à l'aide d'un tableur au profit des outils plus appropriés fournis par R.** Les apprenants y gagneront **en efficacité, en autonomie et en professionnalisme.**

École Pratique des Hautes Études | PSL

# Une organisation à la carte

Pour **satisfaire au plus large éventail de besoins**, la formation se compose de quatre modules indépendants :

1. **Module 1** : Familiarisation avec R
2. **Module 2** : Formatage et nettoyage des données sous R
3. **Module 3** : Visualisation graphique des données sous R
4. **Module 4** : Programmation sous R

**Module 1**
**Familiarisation avec R**

Familliarisation avec l'interface, apprentissage des fonctions de base

**Module 2**
**Formatage de données**

Transformer et nettoyage des données sous R

**Module 3**
**Visualisation de données**

Générer des grahiques de qualité professionnelle sous R

**Module 4**
**Programmer sour R**

Faire réaliser des tâches à R de manière automatisée

## Une pédagogie pratique, en petits effectifs

La formation se focalise sur un apprentissage direct des outils disponibles sous R répondant à un besoin concret de simplicité, d'efficacité et de lisibilité. Grâce à un apprentissage en effectif réduit, l'accent est mis sur la pratique de ces outils, et leur mobilisation par la réalisation d'un projet, possiblement personnalisé, dont la complexité est adaptée au niveau de chaque module.

## Note importante sur l'analyse statistique

Cette formation se concentre sur le travail de nettoyage, mise en forme, visualisation des données à l'aide du logiciel R et ne porte pas sur leur analyse statistique en soi. Pour une formation sur l'analyse statistique des données, voir le **Certificat en analyse de données pour l'écologie et la gestion de la biodiversité.**

# Modalités de la formation

## Inscription

Tout au long de l'année sur envoi d'un CV à **formation.continue@ephe.psl.eu**

## Tarifs

**Plein tarif : 300 € TTC le module.**

**Tarif dégressif : 200 € TTC le module.**

En cas de plusieurs inscriptions par un même employeur ou de réinscription d'une même personne dans un délai de 2 années académiques sur un autre module de la formation R ou pour une personne déjà inscrite dans le cadre de la formation continue sur une autre formation de l'EPHE - PSL.

## Publics concernés

La formation est ouverte à **toutes les disciplines de la biologie** : du domaine de la santé à celui des sciences de l'environnement. Elle s'adresse à toute personne (technicien, ingénieur, gestionnaire, chercheur) voulant monter en compétence sur **la manipulation de données, transformant des heures de manipulations « clique-bouton » rébarbatives sur un tableur en quelques secondes d'exécution** sur R.

# Contacts


## Responsable pédagogique

Pierre de Villemereuil (Maître de conférences, École Pratique des Hautes Études)

**pierre.devillemereuil@ephe.psl.eu**

## Direction de la formation continue

**formation.continue@ephe.psl.eu**


# Objectifs de la formation

La formation vise à ce que les participants puissent :

- **Utiliser le logiciel R** dans une pratique quotidienne ou occasionnelle du traitement de données
- Écrire les **commandes** et un **script** pour traiter des données sous R, à l'aide d'une syntaxe appropriée
- S'orienter vers **les solutions adaptées** aux problèmes qu'ils rencontrent dans la manipulation de leurs données sous R
- **Être autonomes** dans la poursuite de leur apprentissage de l'utilisation de R et de nouveaux outils

# Matériel à prévoir

Un ordinateur portable avec MS Excel ou tout autre logiciel de tableur.

Installer au préalable les logiciels gratuits R et R-studio, téléchargeables à ces adresses (contacter le formateur en cas de difficultés) :

- https://cran.r-project.org/
- https://rstudio.com/products/rstudio/download/

# Programme des modules 1 & 2

NB : Tous les modules se déroulent à Paris (6ème arrondissement)

## Module 1 : Familiarisation avec R

**Pré-requis** Aucun pré-requis pour ce module. Ce module s'adresse à des apprenants n'ayant jamais utilisé R, ou n'ayant utilisé le logiciel qu'au détour d'une formation sans rentrer suffisamment dans les détails pour l'utiliser de leur propre initiative.

**Objectifs** Rendre l'apprenant autonome sous R pour y **effectuer des tâches simples** (importer les données, sélection de colonnes et lignes, calculs simples), et savoir **où trouver les ressources pour progresser** à son rythme.

**Compétences**

- S'approprier le fonctionnement de R, et de l'interface graphique Rstudio.
- Importer des données depuis un fichier de l'ordinateur dans R.
- Utiliser les fonctions de bases de R pour effectuer des manipulations et opérations simples sur les données.
- Trouver en autonomie des informations complémentaires sur une commande R.

**Dates 6 et 7 février 2024**

## Module 2 : Formatage et nettoyage des données sous R

**Pré-requis** Module 1 ou une connaissance de base du fonctionnement de R.

**Objectifs** Donner aux apprenants les clés pour **transformer et nettoyer un jeu de données** chargé sous R, de manière **efficace, automatisée et reproductible.** Cette étape est souvent la plus intimidante et difficile pour un novice, alors que **quelques concepts et fonctions clés** permettent d'effectuer des recherches et transformations complexes à effectuer sous un tableur en quelques commandes. Les outils proposés dans ce module (filtres, outils sur les chaînes de caractère, opérations groupées, pivot, mise en relation de plusieurs jeux de données) visent explicitement à r**emplacer ce travail laborieux sous un tableur par des commandes sous R.** Cela évite les erreurs humaines (copier-coller), mais aussi favorise la reproductibilité (la même commande marchera pour une nouvelle version des données) et la communicabilité (le code suffit à décrire précisément les étapes du « nettoyage » à ses pairs). L'accent sera mis sur les **bonnes pratiques** et comment obtenir des **données bien « rangées »** pour n'importe quel type d'analyse statistique.

**Compétences**

- Nettoyer et filtrer un jeu de données en fonction de critères établis.
- Transformer et formater un jeu de données en vue d'une analyse statistique précise ou pour en produire directement des statistiques résumées.
- Mettre en relation différents jeux de données et en raccorder certaines parties.
- Adopter les bonnes pratiques pour favoriser la reproductibilité et la communicabilité de sa manipulation de données.

**Dates 8 et 9 février 2024**

# Programme des modules 3 & 4

## Module 3 : visualisation graphique des données sous R

**Pré-requis** Module 1 ou une connaissance de base du fonctionnement de R.

**Objectifs** Permettre aux apprenants de **produire des graphiques de qualité professionnelle** en quelques commandes. **La visualisation des données** est un élément extrêmement important pour leur analyse. Elle permet non seulement d'explorer les données pour en saisir les subtilités, mais aussi de transmettre une information quantitative à leur sujet à un public plus large. Choisir **le bon type de graphique**, les bons éléments à y faire figurer et d'autres éléments esthétiques sont cruciaux pour élaborer de bons graphiques. Ce module utilise **ggplot2,** une solution graphique sous R permettant de produire rapidement des graphiques complexes et de qualité professionnelle.

**Compétences**

• Choisir le type de graphique approprié aux données et au message à transmettre.
• Produire un graphique complexe (y compris en combinant plusieurs graphiques) avec le paquet ggplot2.
• Formater un graphique pour une production professionnelle.

**Dates 12 et 13 février 2024**

## Module 4 : programmation sous R

**Pré-requis** Module 2 ou familiarité avec l'utilisation basique de R. Ce module s'adresse à des apprenants ayant suivi les autres modules (au moins le module 2), ou utilisant déjà R, mais voulant approfondir leurs compétences.

**Objectifs** Exploiter R à son **plein potentiel** en automatisant des tâches plus complexes. Certaines tâches sont en effet plus difficiles à automatiser, notamment lorsque les données doivent être analysées ou transformées de manière complexe, ou lorsqu'on veut simuler de nouvelles données. Il est alors nécessaire de découper l'exercice en **un ensemble de tâches plus simples** et de faire appel à **quelques outils plus avancés** de R (boucles, écriture de fonctions, utilisation de listes ou de matrices), y compris en **parallélisation** ces tâches de manière très simple. À l'issue de ce module, les apprenants seront donc capables de **s'attaquer à des problèmes insolubles à l'aide d'un tableur classique** et posséderont l'autonomie suffisante pour **effectuer à peu près n'importe quel traitement de données** (hors analyse statistique) sous R.

**Compétences**

• Découper un problème complexe en une série de tâches plus simples et implémenter ces tâches sous R à l'aide de courtes fonctions.
• Automatiser des tâches répétitives à l'aide de boucles, ou en appliquant des fonctions à des listes.
• Paralléliser très simplement une tâche répétitive à l'aide du paquet future.apply.

**Dates 14 et 15 février 2024**