ENSG Géomatique

ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

Ecole Nationale des Sciences Géographiques

The James Hutton Institute

The James Hutton Institute

Rapport de stage

Cycle des Ingénieurs de l'ENSG 2ème année

Création de cartes environnementales à partir de données satellite

ENSG Géomatique

ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

Charles Laverdure

Août 2021

☒ Non confidentiel ☐ Confidentiel IGN ☐ Confidentiel Industrie ☐ Jusqu'au ...

Ecole Nationale des Sciences Géographiques
6-8 Avenue Blaise Pascal - Cité Descartes - 77420 Champs-sur-Marne
Téléphone 01 64 15 31 00 Télécopie 01 64 15 31 07

Jury

Président de jury :

Pierre-Yves HARDOUIN

Commanditaire :

Alessandro Gimona

Encadrement de stage :

Alessandro Gimona

Enseignant référent :

Cecile Duchêne

Responsable pédagogique du cycle Ingénieur :

Jean-François Hangouët, IGN/ENSG/PEGI

Gestion du stage :

Delphine Genes, Claire Driessens, IGN/ENSG/DFI

Stage de deuxième année du 25/05/2021 au 20/08/2021

Diffusion web : ☒ Internet ☒ Intranet

Situation du document :

Rapport de stage de fin d'année présenté en fin de 2ème année du cycle des Ingénieurs

Nombres de pages : 62 pages dont 24 d'annexes

Système hôte : LaTeX

Modifications :

EDITION	REVISION	DATE	PAGES MODIFIEES
1	0	08/2021	Création

Remerciements

Je tiens tout d'abord à remercier mon tuteur et le responsable de mon accueil au sein de l'institut James Hutton lors de ce stage, Alessandro GIMONA. Merci à Mr. GIMONA pour sa disponibilité, ses conseils et son apport avant et pendant le stage. Dans un second temps, je souhaite remercier Bethany WILKINS, avec qui j'ai travaillé en parallèle sur le projet pour son soutien régulier et les échanges d'informations et de connaissances d'égal à égal que nous avons pu avoir. Je les remercie tout particulièrement pour l'adaptation qu'ils ont pu montrer aux conditions particulières que représente le travail en distanciel.

Je remercie Mme Cecile DUCHENE, mon encadrante au sein de l'ENSG pour la rigueur de son encadrement et la prise de recul que ses conseils ainsi que nos appels ont pu apporter à mon expérience en tant que stagiaire.

Je tiens également à remercier Mme Claire DRIESSENS, qui au-delà de l'aide logistique pour la signature de la convention de stage, m'a fourni le contact de Mr. GIMONA et a appuyé la candidature des élèves de l'ENSG auprès de l'institut James Hutton.

Merci aussi aux élèves de l'ENSG ayant effectué leurs stages au sein de l'institut en même temps que le mien ou lors d'un stage précédent pour l'échange d'information et la vision particulière de leur expérience qui m'ont permis de me décider à choisir ce stage et aussi de le réaliser dans des conditions optimales. Je pense tout particulièrement à Felix QUINTON, élève de l'ENSG promotion 2018, stagiaire de l'institut James Hutton en 2020.

Enfin, je tiens à remercier les autres professeurs de l'ENSG contactés lors de ma prise de décision avant l'acceptation de ce stage pour leurs conseils et leurs retours d'expérience, particulièrement sur l'utilisation du langage de programmation R. Je veux ici citer Mr. Victor COINDET et Mme Malika GRIM en particulier.

Résumé

L'institut James Hutton est une organisation de recherche écossaise reconnue mondialement, dont le principal objectif est l'application de techniques scientifiques afin d'étudier et de donner des directives pour une utilisation durable des ressources naturelles. Lors de ces dernières années, le centre s'est concentré sur la collecte de données concernant la **vulnérabilité sociale et environnementale** des écossais afin de créer des **indicateurs** permettant de décrire les phénomènes d'exposition des populations.

Soucieux de représenter au mieux la présence de **gaz polluants** sur le territoire Ecossais, l'institut se tourne vers de nouveaux jeux de données tels que les relevés de concentrations en gaz dans l'atmosphère par **satellite**. La constellation de satellites du programme **Copernicus** de l'**ESA** permet à travers les données du satellite **SENTINEL 5p** de recueillir de telles informations.

L'objectif de ce stage est donc de créer une série d'**algorithmes** permettant de télécharger les données fournies par les acquisitions satellite et ensuite de les traiter à l'aide de **modèles statistiques** afin d'en retirer des **cartes** de concentration de gaz au-dessus de l'Ecosse.

Cette série d'algorithme est réalisé à l'aide du langage de programmation **R**, un langage adapté à l'analyse statistique et au traitement de données, et est segmentée en plusieurs **fonctions** hautement paramétrables permettant à un utilisateur d'affiner sa sélection en fonction du résultat escompté. Ce programme donne la possibilité d'ajouter d'autres types de données issues du National Atmospheric Emissions Inventory (**NAEI**) ou encore des données altimétriques au travers de modèles numériques de terrain (**MNT**) issus du Service Copernicus de surveillance des terres (**SCST**). Toutes ces données sont agrégées, croisées et traitées grâce à l'utilisation de modèles additifs généralisés (**MAG**).

Le résultat final est donc la production de cartes de concentration de gaz dans l'atmosphère au-dessus de l'Ecosse sur plusieurs intervales de temps, ainsi que des résultats statistiques quant à la **répartition spatio-temporelle** de ces concentrations.

Mots clés : L'institut James Hutton, vulnérabilité sociale et environnementale, indicateurs, gaz polluants, satellite, Copernicus, ESA, SENTINEL 5p, d'algorithmes, modèles statistiques, cartes, R, fonctions, NAEI, MNT, SCST, MAG, répartition spatio-temporelle.

Résumé

The James Hutton institute is a Scottish research organization renown worldwide with the primary objective of applying scientific techniques to study and give directives for a more durable use of our natural resources. For the past few years, the institute has been focusing its research on collecting and using data relating to **social and environmental deprivation** for the Scottish people and creating **indicators** allowing for an accurate description of the exposition phenomenon for the populations.

With the goal of representing as accurately as possible the presence of **polluting gases** over the Scottish territory, the institute started to show interest in new datasets like the ones obtained from atmospheric data retrieval by **satellites**. Satellites part of **ESA**'s **Copernicus** program allow through its mission **SENTINEL 5p** to exploit those types of data.

The main objective of this internship will thus be the creation of a series of **algorithms** that allow for a user to download datasets acquired via satellites and then treat those using **statistical models** with the goal of plotting concentration **maps** for various types of polluting gases over Scotland.

This series of algorithm is to be coded under the programming language **R**, a language that is suited for statistical analysis and data manipulation. This series should also be segmented into multiple **functions** highly configurable that will allow for a user to narrow its selection depending on the required end result. This program also gives the possibility to add other datasets issued by the National Atmospheric Emissions Inventory (**NAEI**) or altimetric data with **DEMs** issued by the Copernicus Land Monitoring Service (**CLMS**). All of these datasets have to be aggregated and then treated using Generalized additive models (**GAMs**).

The end result is the displaying of maps representing the gases concentration especially of pollutants over Scotland for varying time periods as well as statistical results based on the **spatio-temporal repartition** of those concentrations.

Key words : The James Hutton Institute, social and environmental deprivation, indicators, polluting gases, satellite, Copernicus, ESA, SENTINEL 5p, algorithms, statistical models, maps, R, functions, NAEI, DEM, CLMS, GAM, spatio-temporal repartition.

Table des matières

Glossaire et sigles utiles

ENSG École Nationale des Sciences Géographiques
MNT / DEM Modèles numériques de Terrain
NAEI National Atmospheric Emissions Inventory
CLMS Copernicus Land Monitoring Service
MAG / GAM Modèles additifs généralisés
ESA European Space Agency
API Application Programming Interface
HUB Un point de connection

Introduction

La surveillance de la pollution atmosphérique est un concept récent intrinsèquement lié au réchauffement climatique et donc aux notions de développement durable. Dans la lutte généralisée entamée depuis les années 1980-1990 contre le changement climatique et ses effets néfastes sur la santé mais aussi sur les possibilités de développement de l'humanité, la capacité à produire des indicateurs significatifs permettant de qualifier et de quantifier la progression des gaz polluants dans l'air est primordiale. En Ecosse comme partout ailleurs, des scientifiques dont font partie ceux de l'institut James Hutton travaillent dans le but de présenter de manière claire et représentative l'évolution des concentrations de gaz polluants afin d'informer les populations et les autorités.

La surveillance des concentrations de gaz s'est historiquement effectuée au niveau du sol ou grâce à des ballons sondes [1]. Cela dit, devant l'ampleur et l'importance grandissante du problème, l'Agence Spatiale Européenne a lancé fin 2017 la mission SENTINEL 5 précurseur, permettant de faire des mesures atmosphériques par satellite.

L'institut James Hutton effectuait déjà des mesures et analyses de données au niveau du sol mais voulait pouvoir aussi inclure des données du programme Copernicus dans leur modèle. Mon stage constitue la continuation d'une application développée sous R-shiny l'an passé par Nada Boutadghart, stagiaire à l'institut en 2020.

La problématique est donc de savoir comment l'on peut exploiter des données de calculs atmosphériques fournies par l'Agence Spatiale Européenne et sa mission SENTINEL 5p afin de produire des cartes de concentration de gaz polluants ?

Pour ce stage, les principaux objectifs sont la récupération des données, leur analyse et leur ajout dans un modèle permettant de produire des cartes de concentration au-dessus de l'Ecosse.

La présentation de cette étude se divise en trois parties. Dans un premier temps, je présenterai le contexte général de ce stage et sa place dans les travaux plus larges de l'institut James Hutton. Une deuxième partie sera consacrée à la présentation des méthodes d'acquisition des données fournies par la mission Sentinel 5p et des autres acteurs.Nous finirons par présenter l'utilisation des données et l'analyse des résultats.

Chapitre 1
Contexte du stage

1.1 L'institut James Hutton

1.1.1 Présentation et historique

L'institut James Hutton est un centre de recherche basé en Ecosse. Il est reconnu mondialement pour ses activités dans les domaines environnementaux. Les plus de 500 scientifiques du centre apportent une contribution majeure dans la compréhension de grandes problématiques avec un champs d'action à l'échelle globale. Les domaines d'expertise de l'institut se concentrent autour de secteurs clés comme l'énergie, la production agricole ou encore la sécurité alimentaire et environnementale. Etant un centre orienté vers les sujets environnementaux, l'institut James Hutton se consacre aussi à l'étude du changement climatique et à l'impact de l'homme sur son environnement, afin de proposer des solutions pour les populations et aussi les pouvoirs politiques.

Au-delà de ses deux sites en Ecosse (Aberdeen et Dundee), l'institut James Hutton possède de nombreux partenariats à l'international permettant d'étendre son influence et ses capacités d'action. Il emploie des scientifiques ou étudiants internationaux originaires de plus de 29 pays [7] dont je fais partie lors de mon stage. Cela dit, malgré un impact et une portée internationale, le centre reste un organisme public avec un financement encore obtenu à plus de 60% du gouvernement Ecossais [8]. Cette institution est donc résolument ancrée dans l'histoire de la recherche en Ecosse et c'est ce sur quoi nous allons nous pencher maintenant.

L'institut James Hutton est né en 2011 de la fusion entre deux autres centres : le « Macaulay Land Use Research Institute » (**MLURI**) et le « Scottish Crop Research Institute » (**SCRI**)[5] (cf. Annexe A). Le MLURI, fondé en 1930, était spécialisé dans l'application de la recherche comme moyen d'améliorer la productivité des terres écossaises mais aussi dans le partage des territoires entre leurs différents exploitants. Le SCRI de son coté a été fondé sous sa première forme en 1921 et était spécialisé dans la recherche sur la production agricole, notamment la prévention des maladies et pestes en tout genre mais aussi l'impact du changement climatique pour la durabilité, la biodiversité et la qualité de la nourriture produite. Enfin, l'institut regroupant toutes les spécialités évoquées sous la même direction et créé le premier Avril 2011, se prénomme James Hutton d'après une figure du 18ème siècle en Ecosse, consideré par beaucoup comme étant le fondateur de la géologie moderne.

Figure 1.1 – Logo et devise de l'Institut James Hutton

1.1.2 Place des acteurs au sein de l'institut

Comme évoqué précédemment, les domaines de recherche de l'institut sont multiples même s'ils gardent tous une facette environnementale. Pour cela, l'institut James Hutton se divise en 5 départements distincts : Sciences cellulaires et moléculaires, Sciences environnementales et biochimiques, Sciences écologiques, Sciences de l'information et informatiques et Sciences sociales, économiques et géographiques. Mon stage se déroule au sein du département des sciences de l'information et informatiques (ICS [1]). Ce département est dirigé par Rupert HOUGH, un scientifique spécialisé dans la modélisation des risques et dans l'évaluation des menaces environnementales. Ce département couvre un large champ d'expertise allant de l'échelonnage bio-informatique du génome au recueil d'informations climatiques à l'échelle globale [5].

Mon maitre de stage, Alessandro GIMONA est issu de ce même département des sciences de l'information et informatiques. Il est docteur en géographie et écologie spatiale et se spécialise en modélisation mécaniste et statistiques dans ses projets. Il supervise plusieurs projets en parallèle au sein du centre, dont celui qui m'est confié.

Dans un second temps, mon stage est aussi encadré par Bethany WILKINS, une assistante en recherche spécialiste des SIG [2] qui travaille depuis mars 2021 entre autres à la mise en place d'une application utilisant les donnés SENTINEL 5p et que je suis chargé d'aider.

Je me joins donc à cette équipe durant mon stage avec pour but d'aider à la phase de développement de l'application mise en place par Bethany sous la supervision de Mr. GIMONA.

1.2 L'environnement de travail

1.2.1 Organisation et gestion

Etant en année de pandémie, ce stage se réalise pour son intégralité en distanciel. Je reste donc en France, travaillant majoritairement depuis mon domicile ou bien depuis les locaux de l'ENSG lorsque cela est possible. L'institut James Hutton n'a de son côté pas rouvert en capacité d'accueil pré-COVID et donc, dans l'équipe, seul Mr. GIMONA est habilité à travailler depuis les locaux du centre.

Pour la communication entre les différentes branches du groupe, nous privilégions majoritairement des échanges par e-mails mais aussi des réunions sur le logiciel Microsoft Teams. Les échanges d'e-mails se font de manière spontanée dès que le besoin s'en fait ressentir pour échanger sur des problèmes ponctuels pouvant être résolus par écrit. Au-delà de cela, une réunion hebdomadaire tous les mardis matin à 10h (prenant bien évidement en compte le décalage horaire de 1h entre l'Ecosse et la France) permet de cerner les objectifs de la semaine et de discuter des résultats de celle écoulée. Une réunion typique dure entre 30 et 40 minutes, avec une phase de présentation sous la forme d'un partage d'écran de ma part suivi d'une discussion qui permet d'amener vers les objectifs pour le reste du projet. Lors de ces réunions, les échanges se font principalement avec Mr. GIMONA qui supervise le projet avec plus de recul et est plus à même de guider les réflexions ou de répondre aux questions sur des points de blocage. Les échanges avec Bethany WILKINS via Teams sont quant à eux plus fréquents et moins formels, au début pour me permettre de se remettre à son niveau étant donné qu'elle travaille sur le projet depuis mars, et à la fin du stage afin de valider et tester les parties de codes que j'ai produites.

1. Information and Computational Sciences
2. Système d'Informations Géographiques

En dehors de ces réunions, le travail s'effectue avec une très grande autonomie, les directives hebdomadaires sont larges et le sujet est orienté comme un véritable sujet de recherche qui me laisse la possibilité d'explorer plusieurs pistes et de faire de nouvelles propositions en cours d'avancement. Je jouis d'une grande liberté de travail et chaque nouvelle piste est débattue lors de nos réunions hebdomadaires. Cette grande liberté et une définition de sujet laissant de la place à l'exploration et parfois même l'interprétation à rendu les outils de gestions essentiels. Un diagramme de GANTT prévisionnel [voir Annexe B] a été produit grâce au logiciel EXCEL à la fin de la période d'analyse et a été révisé à de multiples reprises avant d'arriver à sa version finale que vous pouvez consulter ci-dessous (1.2). Comme vous pouvez le constater, ce diagramme a été très amandé, particulièrement au moment nous avons décidé de ne plus fournir une application réalisée sous R-shiny mais simplement un enchainement de fonctions. (cf. partie 1.3.2).

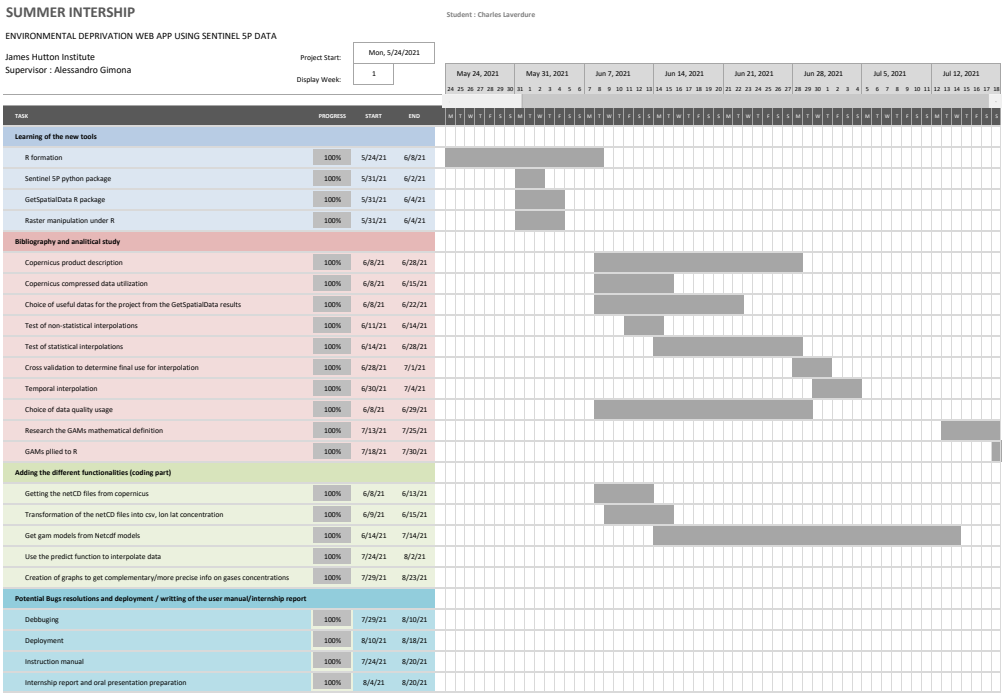


FIGURE 1.2 – Diagramme de GANTT révisé du 11 août 2021, disponible en taille entière en Annexe B

La majorité du stage reposant sur du développement de code informatique, nous utilisons la plateforme GITHUB afin de s'échanger nos fichiers mais aussi pour garder une copie numérique de nos codes en cas de problème en local sur un ordinateur. Sur ce GITHUB partagé, j'effectue des dépôts hebdomadaires (souvent le vendredi) qui permettent aux autres membres du groupe de tester les fonctions et aussi de faire leurs retours ou bien d'aider en cas de nécessité de débogage.

Enfin, tout au long de ce stage j'entretiens un carnet de suivi réalisé sous LaTeXqui permet de relater des avancements mais aussi des problèmes de chaque semaine. Ce carnet se révèle très utile au moment de la rédaction de la documentation utilisateur pour le code en fin de stage mais aussi pour la rédaction de ce rapport.

1.2.2 Outils utilisés

L'une des seules contraintes inamovibles de stage est que l'ensemble du code soit réalisé sous le langage R. il s'agit d'un langage adapté à la gestion de données et surtout optimisé pour les sciences statistiques [11]. Ce langage est disponible sur tous les systèmes d'exploitation et dispose d'une large communauté qui permet de trouver facilement de la documentation ou de se renseigner sur des forums d'entraide.

Je ne disposais d'aucune formation à cet outil et ai donc dû prendre deux semaines au début du stage afin de me familiariser avec ce nouveau langage qui ressemble cependant beaucoup au langage Python que je maitrise déjà. Mon maitre de stage m'a fourni plusieurs cours en ligne centrés sur les thématiques lies au sujet du stage que j'ai pu suivre en autonomie.

Durant cette période de formation, j'ai pu tout particulièrement découvrir le concept de *packages* sous R. Un *package* représente un module, un ensemble de fonctions permettant de réaliser des opérations spécifiques à un certain champs d'application. Lors de ce stage, les *packages* s'étant révélés les plus utiles sont ceux orientés vers l'analyse géospatiale. Je vais ici en décrire tout particulièrement deux qui se retrouvent dans plusieurs parties du programme.

Tout d'abord le *package* **raster**. Ce module permet l'importation, l'exportation et la modification de données au format raster. Il est très utile dans le contexte de ce stage car le résultat final attendu est une carte de concentration en gaz polluants qui se présente donc sous la forme d'une grille dont essayera de maximiser la résolution spatiale (ie. réduire au maximum la taille de chaque cellule afin d'obtenir des données locales). Dans un second temps, le *package* **sf** qui constitue en fait un rassemblement de trois sous modules : **rgdal**, **sp** et **rgeos**. **Rgdal** permet tout comme le *package* **raster** des manipulations de données raster mais avec des opérations pouvant être optimisées. Le module **sp** permet quant à lui de définir au sein du code sous R des types de données spatiales telles que les SpatialPoints (des points), les SpatialLines (des lignes) et les SpatialPolygons (des polygones) qui permettent ensuite de manipuler des éléments spatiaux grâce au module **rgeos**.

On peut finir par présenter le *package* **getSpatialData** [13]. Ce module constitue le point de départ du projet car il permet d'interagir avec l'API [3] du HUB Copernicus afin de télécharger des données issues des missions SENTINEL. Ce package est développé dans un esprit d'open source par J. SCHWALB-WILLMANN, un scientifique de l'université de Würzburg en Allemagne. Grâce à ce module, on peut mettre en place le filtrage et un début d'analyse des données de concentration en gaz dans l'atmosphère fourni par la mission SENTINEL 5p.

Maintenant que nous avons présenté le langage R, voyons comment il s'utilise. Comme tous les langages, R ne peut pas tourner tout seul, pour qu'il y ait une interaction entre l'homme et la machine, nous avons recours à un environnement de développement intégré (IDE [4]). Il existe de nombreux IDE permettant d'exécuter une série d'instructions codées en R. Mon choix s'est porté sur l'IDE **Rstudio** car c'est l'environnement le plus largement utilisé à travers la communauté, mais aussi l'un des plus simple à prendre en main lorsque l'on débute. Il permet notamment une gestion automatique des packages, une auto-complétion des lignes de commande et comprend aussi des systèmes de contrôle de version dans le cas de partage de code.

Pour ma part, suite à un problème lors de l'installation d'un package, j'utilise la version émulée de **Rstudio** via la plateforme Anaconda. Cela-dit, l'interface reste la même que sur une version Desktop classique.

3. application programming interface
4. Integrated Development Environment

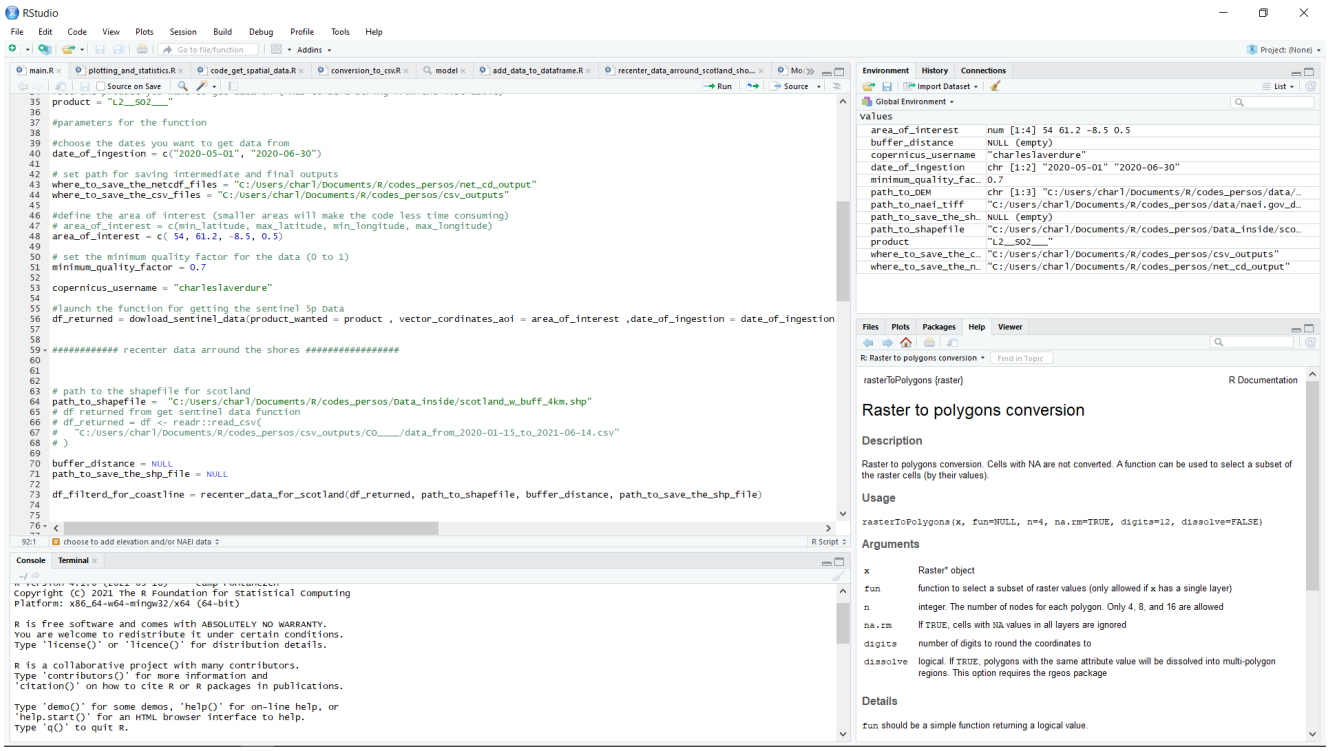


FIGURE 1.3 – Screenshot de l'IDE R-studio

L'interface utilisateur de **Rstudio** montrée ci-dessus inclue 4 parties :

— **La fenêtre d'édition** : Cette fenêtre permet d'écrire du code avant de l'exécuter, soit sur l'entièreté d'un fichier soit en envoyant les instructions ligne par ligne vers la console.
— **La console** : La console est l'espace où le code est effectivement exécuté.
— **La fenêtre workspace et d'historique** : Cette fenêtre permet de retrouver les variables actives dans l'environnement mais aussi un historique des dernières opérations effectuées.
— **La fenêtre pour les fichiers/graphiques/packages/aide** : Cette fenêtre est plus générale et permet comme son nom l'indique d'afficher un explorateur de fichier, d'afficher des graphiques, de présenter l'ensemble des packages disponibles ou bien encore de consulter l'aide fournie par R pour chaque fonction.

1.3 Les objectifs du projet

1.3.1 Travaux précédents et continuation

Cette partie vise à présenter les travaux préliminaires ou complémentaires à mon stage. Elle se divise en deux blocs : le travail de développement d'une application permettant de visualiser des données liées aux vulnérabilités sociales et environnementales de la population écossaise effectué par Nada BOUTADGHART, et aussi le travail d'analyse et de pré-traitement effectué par Bethany WILKINS en charge du projet dans lequel s'intègre mon stage.

Lors de son stage l'année passée, Nada BOUTADGHART a effectué un stage au sein du département ICS tout comme moi et aussi sous la supervision de Mr. GIMONA. Le but de son projet était de créer une application sous R-shiny, une extension de R qui donne la possibilité de réali-

ser des applications paramétrables pour le web, permettant à un utilisateur de se renseigner sur la situation environnementale et sociale en Ecosse. Cette application rassemble des données majoritairement fournies par le SIMD (Scottish Index of Multiple Deprivation) et permet de réaliser des cartes qui agrégent ces données. Les données SIMD sont un croisement entre des données de relevé de pollution au niveau du sol avec des données sociaux-économiques sur la population Écossaise.

Bethany WILKINS quant à elle est une employée récemment engagée par l'institut James Hutton. Depuis mars 2021, Bethany travaille sur l'analyse de données issues du hub Copernicus. Avant mon arrivée au sein de l'institut, Bethany a effectué un travail d'analyse et de recherche qui a amené à la découverte notamment de la librairie GetSpatialData dont nous avons discuté précédemment. De plus, elle a réalisé une mise au point des objectifs de la mission avec en point de mire la réalisation d'une application similaire à celle de Nada BOUTADGHART via le module R-shiny incluant des données issues de satellite. J'ai donc été recruté au sein de l'institut afin d'assister Bethany dans la réalisation de cette application. Son travail préliminaire s'est révélé très utile pour mon stage car il a permis de recentrer les recherches de librairies R.

1.3.2 Objectifs correspondants au stage

Pour en venir aux objectifs de mon stage, ils m'ont été présentés en début de stage comme la « continuation d'une application R-shiny déjà existante et permettant à un utilisateur de retirer des informations sur la vulnérabilité sociale et environnementale en utilisant des données SENTINEL 5p pour toute l'Ecosse ou plus localement avec une possibilité d'obtenir des moyennes hebdomadaires, mensuelles ou annuelles ». Comme on le remarque, cette énonciation est très large et peut porter à interprétation, il faut donc procéder à une reformulation du sujet et à une redéfinition complète en fin de stage. Il faut distinguer deux parties principales prévues dans le stage initialement, une partie de traitement de données sous la forme de fonctions R et une autre partie de mise en place d'une application paramétrable et adaptée à des utilisateurs grand public. Les objectifs de début de stage sont donc comme suit :

- Se former à l'outil de programmation R.
- Réaliser une phase d'analyse permettant de distinguer les données nous intéressant et à ajouter à l'application.
- Mettre en place une fonction permettant de télécharger des données issues de la mission SENTINEL 5p. Ces téléchargements doivent être paramétrables pour obtenir des données entre certaines dates et aussi pour des régions spécifiques.
- Obtenir d'autres données à croiser avec les données SENTINEL 5p.
- Chercher un moyen d'intégrer ces nouvelles données d'une façon significative qui puisse mettre en valeur les concentrations issues du HUB Copernicus.
- Interpoler les données afin de produire des cartes de concentration uniformes sur l'ensemble du territoire Ecossais exportables sous un format raster.
- Extraire des statistiques à partir des données raster afin de matérialiser des tendances d'évolution temporelles des phénomènes en des points géographiques précis.
- Mettre en place ces différents codes sous la forme d'une application web via le module R-shiny.

A la fin de ce stage, la plupart des objectifs ont pu être accomplis, cependant d'autres ont été amandés. Devant les temps de calculs importants liés aux différentes opérations géospatiales sous R, l'idée de réaliser une application WEB a finalement été abandonnée. Cela dit, l'objectif n'est pas

complètement abandonné pour la suite de la réalisation du projet et donc le choix à été fait de créer toutes les fonctions permettant de réaliser les opérations avec un grand degré de paramétrisation. Ce degré de paramétrage élevé du code, même si toujours sous la forme d'une série de fonctions manuscrites, permettra dans le futur une intégration facilitée sous R-shiny si des moyens sont trouvés pour optimiser fortement le code ou alors, comme évoqué avec le maître de stage, si l'application est mise en place sur des supercalculateurs dans le réseau interne de l'institut James Hutton.

Le programme mis en place en fin de stage permet à un utilisateur de télécharger des données SENTINEL 5p de colonnes de gaz allant de la surface du sol jusqu'aux limites de la troposphère. Une fois ces données chargées, le choix est laissé à l'utilisateur d'ajouter des données altimétriques ou de concentration au niveau du sol. Afin d'agréger toutes ces données et dans le but d'une interpolation menant à la réalisation de cartes, le choix se porte sur l'utilisation de modèles additifs généralisés. Ces modèles, relevant des sciences statistiques et proposés par Mr. GIMONA permettent d'extraire les relations entre les différents paramètres. Une fois calculé, l'utilisateur peut choisir de produire des cartes de concentration à la suite d'une interpolation.

On peut donc considérer que la majorité des objectifs du stage ont été remplis, mis à part la partie de calculs statistiques pixel par pixel et aussi la mise en place sous la forme d'une application WEB sous R-shiny. Nous allons maintenant présenter la démarche qui a permis d'arriver à ses résultats.

Méthodes d'acquisition de données

Chapitre 2

2.1 Le programme SENTINEL 5p – TROPOMI

2.1.1 Présentation des outils

Le programme Copernicus SENTINEL-5 précurseur est le premier programme de l'agence spatiale européenne (ESA) entièrement destiné à l'observation de l'atmosphère [13]. Il a été lancé en 2017 avec l'envoi du satellite du même nom en orbite. Cet achèvement résulte de la collaboration entre une multitude d'acteurs dont l'ESA, la Commission Européenne, le centre spatial Néerlandais ainsi que des industriels et des scientifiques. Il permet une couverture spatio-temporelle élevée avec des passages réguliers autour de l'ensemble du globe. Au-delà des acquisitions satellites brutes qui permettent de déceler les concentrations en gaz dans l'atmosphère grâce à la diffraction des rayons électromagnétiques, le programme SENTINEL 5p s'accompagne d'une multitude de post traitements qui rendent les données utilisables par le grand public dans un esprit d'Open-Source.

Le type de données fournies par le programme varient et vont des relevés de concentrations en gaz sous la forme de colonnes à l'observation des nuages en passant par la mesure des niveaux d'UV (voir Annexe C pour une description complète des produits).

Dans notre cas d'utilisation, nous utilisons des données de concentration sous la forme de colonnes verticales ayant une résolution spatiale allant jusque 5.5 km x 3.5 km [14]. Ces colonnes ne sont disponibles dans leur entièreté que pour 6 produits, 6 polluants majeurs présents dans notre atmosphère (cf. 2.1.1). Ces colonnes représentent en mol par m2 la concentration totale en gaz entre la surface du sol et le satellite.

Tableau des gaz disposant d'un attribut "colonne totale"		
Nom du produit	Abréviation	Nom SENTINEL
Monoxyde de carbon	CO	L2__CO_____
Ozone	O	L2__O3_____
Méthane	CH_4	L2__CH4____
Formaldehyde	HCHO	L2__HCHO___
Dyoxide de Nitrogen	NO_2	L2__NO2____
Dioxide de sulfure	SO_2	L2__SO2____

Afin de télécharger ces données, nous avons recours à la librairie R GetSpatialData qui permet de transférer les données du HUB Copernicus vers le logiciel R pour poursuivre les traitements. Pour une utilisation classique du module, l'utilisateur doit renseigner ses identifiants personnels vers le HUB. La création, tout comme l'utilisation des données est complètement gratuite et ouverte à tous. Ensuite, l'utilisateur choisit une période et le nom de la mission à partir de laquelle il souhaite récupérer les données. Dans notre cas, le nom de cette mission est bien évidemment « sentinel-5p » et l'on peut insérer toutes les dates comprises entre la mise en ligne du programme en 2017 et la date du jour. A ce propos, la mission SENTINEL-5p propose 2 types de données en fonction du nombre de traitements effectués dessus. La première catégorie se nomme NRTI (near real time) et permet, comme son nom l'indique, d'obtenir des données en quasi temps réel après l'acquisition. La seconde catégorie est OFFL (offline) et met un peu plus de temps après l'acquisition pour être disponible au grand public car des traitements sont effectués sur les données brutes pour en extraire plus d'indicateurs. La notion d'instantanéité n'étant pas une priorité du projet, nous avons choisi les données plus stables fournies par le jeu de données OFFL.

Le dernier paramètre à rentrer dans la fonction est bien évidemment l'emprise souhaitée. Pour cela, on définit une aire d'intérêt comprenant la totalité de l'Ecosse sous la forme d'un rectangle renseigné par les coordonnées de ses quatre extrémités. Enfin, avant de lancer le téléchargement, l'utilisateur choisit les données du gaz polluant qu'il désire.

Une fois tous les paramètres de téléchargements entrés dans la fonction, le module se connecte à l'API du HUB Copernicus et lance le téléchargement des données. Ces données sont alors reçues sur l'ordinateur de l'utilisateur dans un format bien particulier contenant beaucoup d'informations, les fichiers Netcdfs. Nous allons discuter de la composition et de l'utilisation de ces fichiers dans la prochaine partie.

2.1.2 Organisation des données

Les données chargées à partir de l'API du HUB Copernicus se présentent sous la forme de fichiers Netcdf. Les fichiers Netcdf [1] sont des fichiers compressés qui regroupent un grand nombre de données de formats et de types différents. Dans le cas des fichiers fournis par la mission SENTINEL 5p, ils se décomposent en de grandes catégories imbriquées, allant des données traitées au métadatas en passant par les données brutes (cf. 2.1).

Pour chaque passage d'acquisition du satellite, un fichier est produit avec un volume très large de données. Toutes les données présentes dans ces fichiers n'ont pas une forte importance par rapport à l'utilisation finale que nous recherchons. En moyenne, chaque fichier est d'une taille de 500Mo, soit approximativement 1.5G de donnée par jour d'acquisition. Traiter des fichiers de cette taille peut se révéler très rapidement impossible surtout si notre objectif est d'étudier des tendances hebdomadaires ou mensuelles.

La solution est donc de décompresser chacun des fichiers afin d'en conserver uniquement les données pouvant nous intéresser. Des librairies R telle que netcdf4 permettent ce type d'opération. Une fois ces fichiers décompressés, on peut en afficher le contenu exploitable (voir Annexe D pour un exemple de liste complète)

Une phase intensive d'étude et d'analyse est nécessaire afin de pouvoir naviguer à travers ces fichiers et d'en comprendre les composants. Le site de l'ESA [4] fournit des manuels d'utilisation très détaillés pour chaque produit ce qui permet d'identifier les éléments dont nous avons besoin.

1. Network Common Data Form

De ces fichiers, nous devons retirer en priorité les données spatiales et temporelles des concentrations en gaz dans les colonnes atmosphériques. Les 4 éléments que nous choisissons donc d'extraire sont, pour chaque point : longitude, latitude, temps en UTC [2] et la valeur de la concentration. Chaque point représente une acquisition c'est-à-dire un rayon électromagnétique envoyé vers la surface de la terre et retourné vers le satellite dont on retire dans notre cas l'horaire exacte d'émission, les coordonnées de l'impact à la surface terrestre et la valeur post-traitement.

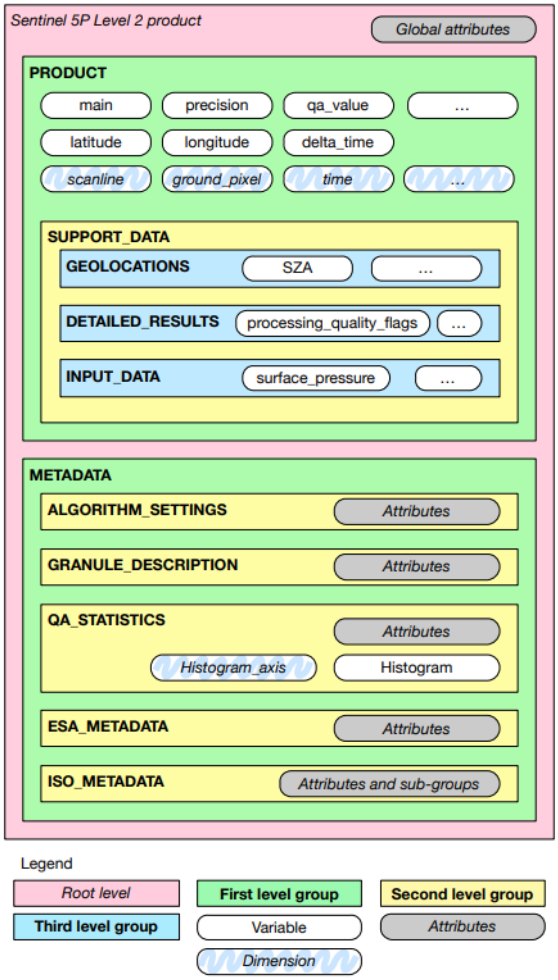


FIGURE 2.1 – Diagramme de composition d'un fichier Netcdf comme présenté dans un manuel utilisateur d'un produit SENTINEL 5p

D'autres éléments peuvent être intéressants pour notre projet mais n'on pas été retenus dans un souci de simplification et d'optimisation de l'application, leur utilisation possible sera abordée dans la partie discussion et continuation à la fin de ce rapport. Nous conservons donc ces quatre données, deux spatiales, une temporelle et une de concentration avant d'y appliquer d'autres filtres qui seront expliqués ci-après.

2. Universal Time Coordinated

2.2 Autres données utiles au projet

2.2.1 Données altimétriques

Les données de concentration en gaz fournies par SENTINEL 5p sont sous la forme de colonnes allant de la surface du sol jusqu'à la troposphère. On se rend donc compte que la valeur des données recueillies est très dépendante de l'altitude du terrain. La surface du sol n'étant pas régulière en Ecosse il s'impose de faire intervenir des données altimétriques avant de calculer un modèle permettant une interpolation finale.

Le jeu de données fourni par le service Copernicus de surveillance des territoire [3] utilisé dans ce projet est un modèle numérique de terrain qui a été recoupé pour couvrir les frontières administratives de l'Ecosse (cf. 2.2). Il s'agit d'un MNT publié en 2016 et qui a pour caractéristique une résolution spatiale de 25m avec une précision verticale de plus ou moins 7 mètres.

FIGURE 2.2 – MNT du territoire Ecossais issus du CLMS

A partir de ce MNT et grâce à la fonction extract du package raster, nous donnons pour chaque point extrait de la mission SENTINEL 5p une valeur d'altitude. Cette valeur d'altitude est conservée telle quelle et sera utile dans le calcul du modèle additif généralisé présenté en partie 3.1

2.2.2 Données de concentration au niveau sol

L'ajout de données de concentration en gaz dans le programme a d'abord été imaginé dans un but de vérification et de croisement avec les données issues du satellite. Cependant, quand l'inclusion de modèle additif généralisé a été proposé, cette inclusion pouvait se faire de manière plus directe

comme l'un des paramètres appartenant au calcul du modèle.

Les données de concentration au niveau du sol sont extraites du site du National Atmospheric Emission Inventory. Il s'agit d'un institut publique britannique fondé en 1970 avec pour but le recueil d'informations et de données sur les gaz polluants présents sur le territoire de la Grande-Bretagne [10]. Les données fournies par l'organisation sont libres de droits et accessibles à tous.

Ces données se présentent sous la forme de fichiers raster et sont particulièrement intéressantes car elles classent les rejets de gaz dans l'atmosphère par type d'émetteur. Le seul problème que nous rencontrons est que ces données ne sont disponibles que comme des moyennes annuelles et seule la dernière version éditée est disponible à chaque nouvelle édition. Ainsi, nous avons travaillé avec des données concentrant des relevés pour l'année 2018 mais ne pouvions pas effectuer de comparaisons d'une année sur l'autre avec ce type de données. Cela dit, ces données se révèlent tout de même très utiles pour identifier les centres de faibles et fortes émissions et permettent donc de valider en partie, à l'échelle macroscopique par rapport au données satellites, les données issues de SENTINEL 5p.

Enfin, une autre limite dans l'exploitation de ces données repose dans le fait que tous les produits proposés par la mission SENTINEL 5p ne sont pas surveillés par le NAEI. Ainsi seul 3 des 6 gaz polluants extraits des données satellites peuvent être croisés avec des données de concentration sol : le monoxyde de carbone, le dioxyde de souffre et le méthane.

Nous allons maintenant discuter comment filtrer toute ces données pour ne conserver que celles qui apportent un effet significatif au projet.

2.3 Filtrage et agrégation

2.3.1 Choix utilisateur pour le filtrage et le type de données

Nous avons vu lors des précédents paragraphes que les jeux de données recueillies comportent un grand nombre d'informations et sont donc trop larges pour être traitées efficacement avec même le risque d'inclure des données parasites. Il est donc essentiel de filtrer les différentes données acquises avant de procéder aux traitements.

Dans un premier temps, le moyen le plus simple de réduire le nombre de points présents dans notre acquisition vers un jeu de données utiles est de recentrer les informations sur la partie terrestre de l'Ecosse. On cherche en effet à mesurer la vulnérabilité environnementale des populations et donc les surfaces des mers ne sont pas importantes. Pour cela, on réalise un simple recadrage des points se situant dans les limites administratives du territoire Ecossais. (cf. fig. 2.3 ci-dessous).

Dans un second temps, nous nous servons d'un des indicateurs fournis par la mission SENTINEL 5p afin de garder les points les plus intéressants. L'un des paramètres extractibles depuis les fichiers Netcdf se nomme « qa_value ». Ce paramètre agrège un ensemble d'autres indicateurs nommés « flags » qui permettent de juger de la qualité d'une acquisition. La « qa_value » est un chiffre compris entre 0 et 1, 0 signifiant qu'aucune donnée n'a pu être retirée pour un point et 1 signifiant une extraction parfaite. La qualité d'un relevé peut être influencée par de nombreux paramètres. En premier lieux, la réflectance de la surface au sol influe sur le retour que le satellite peut obtenir. Ainsi des surfaces enneigées ou avec un albedo élevé en général auront une qualité moindre. Le deuxième facteur influant la qualité d'une acquisition est la présence de nuages. Les rayons électromagnétiques envoyés par le satellite sont en effet bloqués par les nuages et ainsi, pour avoir des résultats de qualité, la zone d'étude doit être dégagée. Enfin, le dernier paramètre majeur influent la qualité d'une acquisition est

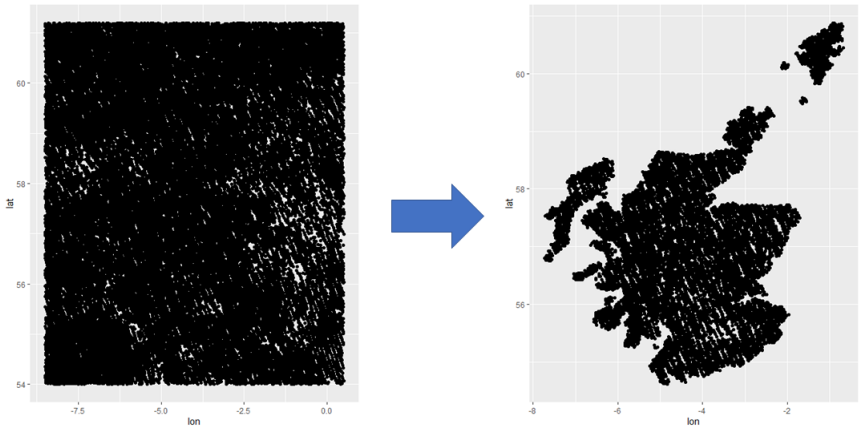


FIGURE 2.3 – Processus d'extraction des points au à l'intérieur des terres

le post-traitement de la donnée. Si, quand les données sont traitées par le logiciel de post-traitement de l'ESA, un calcul sort de la norme ou demande des informations qui sont manquantes, la valeur de l'indicateur « qa_value » est revue à la baisse.

Dans le programme, le choix est laissé à l'utilisateur de déterminer la valeur minimum de qualité des acquisitions à retenir. Une valeur générique fournie par la mission SENTINEL 5p est disponible pour chaque produit en fonction des applications recherchées dans le manuel d'utilisation correspondant au produit.

Enfin, la dernière possibilité de filtrage concerne les données fournies par le NAEI. En effet, dans le processus d'extraction de données expliqué dans le paragraphe précédent, certaines valeurs ne peuvent pas être extrapolées et l'on se retrouve donc avec des points sans valeur de concentration pour certains types d'émetteurs. Le choix est donc laissé à l'utilisateur du pourcentage minimum de valeurs présentes pour chaque type d'émetteur, sinon la colonne correspondante n'est pas retenue.

2.3.2 Réalisation et résultats

Une fois toutes ces données acquises et filtrées, le résultat se présente sous la forme d'un tableau ou « dataframe » dans le langage R. Ce tableau (dont l'en-tête est présenté ci-dessous cf.fig 2.4) contient un nombre de colonnes correspondant aux nombres de paramètres et de données que l'utilisateur souhaite incorporer. Ce tableau comprend au minimum 7 colonnes correspondant aux données recueillies via le satellite SENTINEL 5. A ces 7 colonnes peut venir s'ajouter une colonne correspondant à l'altitude du point et enfin, une colonne est rajoutée par type d'émetteur fourni par le jeu de données du NAEI.

Chaque ligne correspond donc à une acquisition, à un point précis dans l'espace et le temps.

Filter

	X1	lat	lon	pr_tc	qa_tc	time	ndate	elevation	domcomco18	energyprodco18	indcomco18	indprocco18	natureco1
1	1	56.56140	-2.600715	2.002972e+18	1.0	2020-01-15 11:13:27	1579086807	27.760098	9.96335602	0.0014177550	0.125842452	NA	0.25173
2	2	56.60106	-2.657056	1.959425e+18	1.0	2020-01-15 11:13:28	1579086808	73.616226	0.39046818	NA	NA	NA	0.06294
3	3	56.82325	-2.286264	1.810619e+18	0.7	2020-01-15 11:13:28	1579086808	NA	9.55226040	0.0005592782	NA	NA	0.07047
4	4	56.64087	-2.713317	2.024686e+18	0.7	2020-01-15 11:13:28	1579086809	60.258846	0.60095322	0.0044502094	0.001614454	NA	0.18173
5	5	56.86323	-2.342496	1.814235e+18	0.7	2020-01-15 11:13:28	1579086809	59.012314	0.26720756	0.0114831347	NA	NA	0.37256
6	6	56.68041	-2.770007	1.935092e+18	0.7	2020-01-15 11:13:29	1579086810	143.570877	0.25540391	NA	NA	NA	0.05282
7	7	56.90296	-2.399145	1.818663e+18	0.7	2020-01-15 11:13:29	1579086810	110.632179	0.65814966	0.0008424987	NA	NA	0.07338
8	8	56.71999	-2.826713	1.932055e+18	0.7	2020-01-15 11:13:30	1579086811	85.142113	0.25168306	0.0038225122	NA	NA	0.16476
9	9	56.94272	-2.455814	1.799038e+18	0.7	2020-01-15 11:13:30	1579086811	256.586975	NA	0.0193447080	NA	NA	0.58620

FIGURE 2.4 – Début du tableau avec toutes les données possibles ajoutées

Chapitre 3
Analyse des données récupérées & résultats

3.1 Analyse spatio-temporelle par GAM

3.1.1 Les GAMs appliqués au projet

Les modèles additifs généralisés (GAMs)[1] [12] représentent un champ d'application des statistiques relativement récent permettant de représenter des interactions complexes entre plusieurs paramètres. Ils se situent à mi-chemin entre les modèles d'analyse statistiques linéaires et le machine learning. Ils permettent donc de calculer et faire ressortir des tendances entre des paramètres avec une relation trop complexe pour être simplement représentés par un modèle linéaire.

Nous utilisons les GAMs au sein du projet dans un but d'interpoler nos données de concentration sur une grille régulière au-dessus de l'Ecosse. Ce processus d'interpolation sera détaillé dans la partie 3.4. En attendant, attardons-nous sur le processus de modélisation via GAM.

Sur la figure 3.1 illustrée ci-dessous, nous pouvons voir la relation complexe entre deux différentes variables x et y. la figure 3.2 illustre l'interprétation de ces données par un modèle linéaire sur la gauche et l'on voit donc que celui-ci n'est pas adapté pour représenter notre jeu de données. Enfin, sur la gauche, on peut voir comment un modèle GAM permet d'ajuster notre prédiction au jeu de données avec précision.

D'un point de vu mathématique, nous passons d'une équation de la forme :

$$y = \beta 0 + x1\beta 1 + \epsilon, \epsilon \tilde{N}(0, \sigma 2) \tag{3.1}$$

Pour une équation linéaire a une équation où un terme de lissage est ajouté :

$$y = x1 + f(x1) + \epsilon, \epsilon \tilde{N}(0, \sigma 2) \tag{3.2}$$

Cela dit, un modèle de GAM peut être une composition de termes linéaires et de lissage comme dans l'exemple suivant :

$$y = \beta 0 + x1\beta 1 + f(x2) + \epsilon, \epsilon \tilde{N}(0, \sigma 2) \tag{3.3}$$

1. General additive models

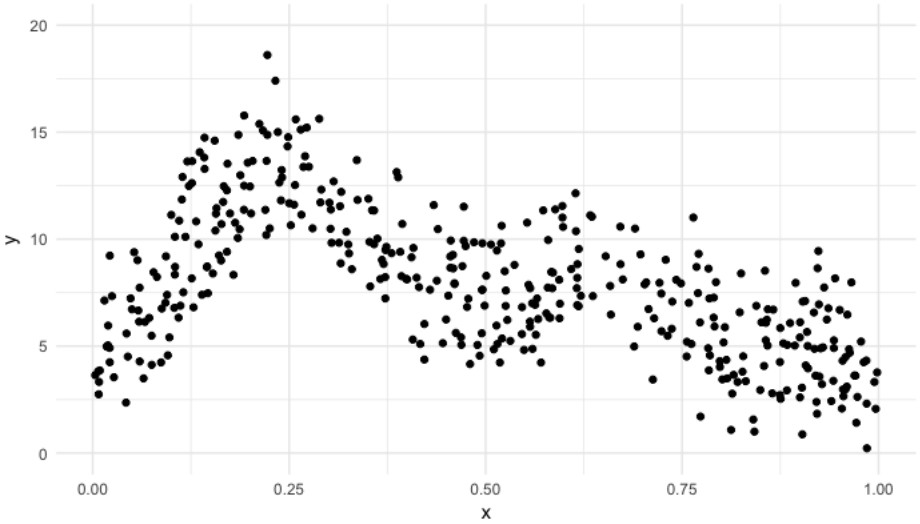


FIGURE 3.1 – Plot test pour 2 données x et y.
Source des visuels : [12]

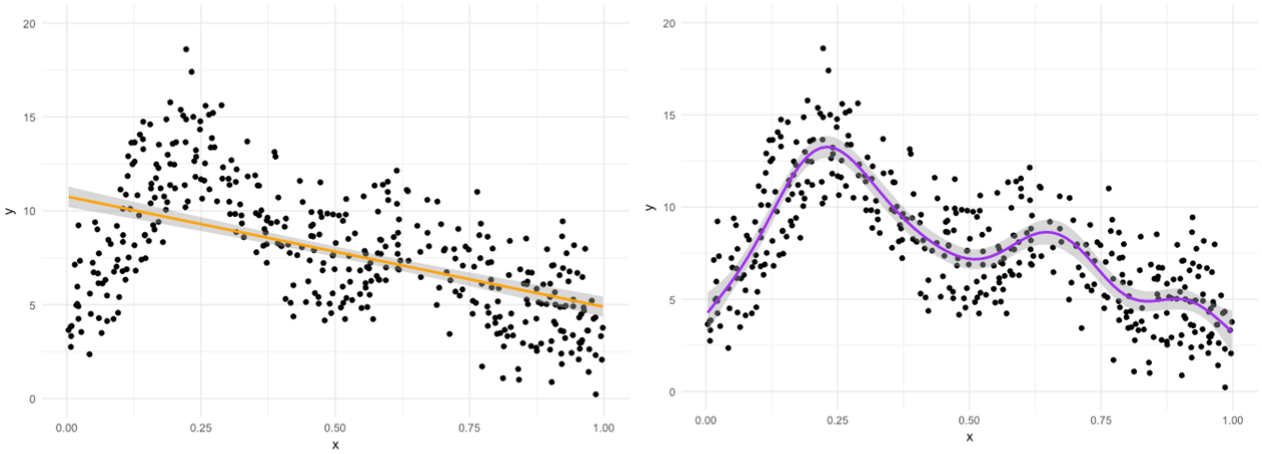


FIGURE 3.2 – Plot test pour 2 données x et y.
Source des visuels : [12]

Un modèle GAM est en fait un assemblage de multiples fonctions dites basiques qui permettent de représenter une fois combinées des interactions complexes (cf. fig. 3.3).

Enfin, pour finir de décrire les possibilités des GAMs, nous devons parler de la possibilité d'ajouter des tenseurs à l'intérieur des modèles. Les tenseurs représentent l'interaction entre deux paramètres à des échelles différentes comme l'espace et le temps par exemple. Ils permettent de représenter de façon continue ces deux paramètres dans le même terme de lissage. Cela résulte en une équation du type suivant :

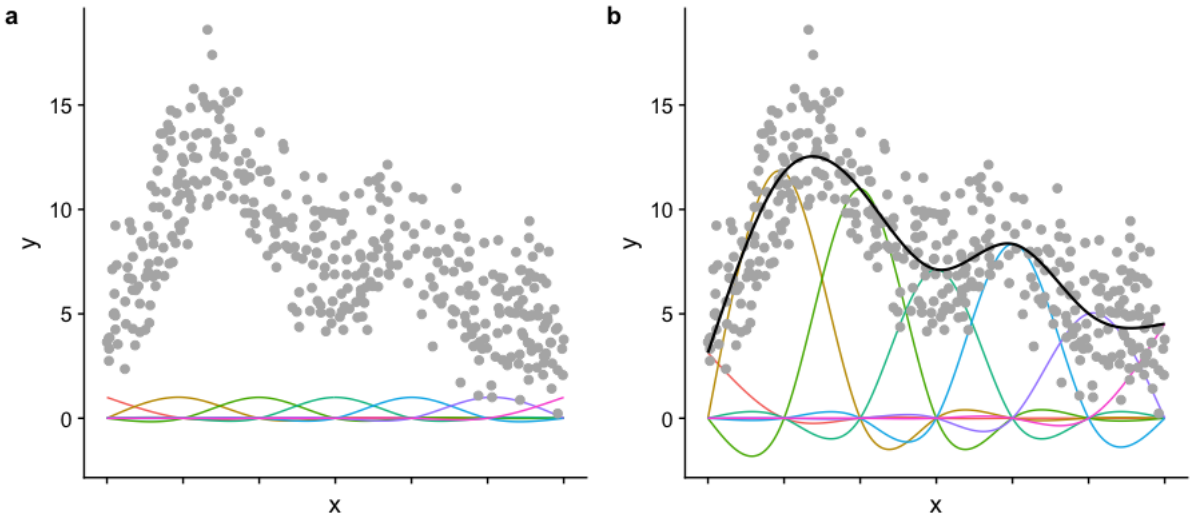


FIGURE 3.3 – Graphique montrant la composition d'un ajustement de donnée par GAM.
Source des visuels : [12]

$$y = \beta 0 + x1\beta 1 + f(x2) + ti(x1, x2) + \epsilon, \epsilon \tilde{N}(0, \sigma 2) \tag{3.4}$$

Toutes ces formules seront précisées plus en détail dans la partie suivante grâce à des exemples appliqués au données du projet.

Le choix de modèle GAM a aussi été effectué car un grand nombre de références et d'articles s'appliquant à des données aux relations complexes et de type spatio-temporelles comme celles que nous traitons au sein du projet. Je me suis particulièrement appuyé sur les travaux de Wikle, Zammit-Mangion et Cressie [2] mais aussi de Mitchell Lyons [9] pour leur travaux sur les GAMs appliqués aux géosciences.

3.1.2 Apport des différents types de données

Pour rappel, nous disposons de quatre types de données principales dont nous voulons étudier l'impact sur la concentration issue des données de la mission SENTINEL 5p. Ces quatre types de données sont : un paramètre de temps, issu des données SENTINEL et donnant l'horaire exact de l'acquisition. Le deuxième type de paramètre est spatial, il s'agit des valeurs de latitude et longitude de chaque point d'acquisition. Le troisième type de donnée sont les données altimétriques qui donnent l'altitude de chaque point du jeu de données. Enfin le dernier type de données représente la concentration moyennée annuelle en 2018 par gaz fournie par le NAEI.

On remarque que ces quatre types de données sont très différentes les uns des autres, que ce soit sur un terme d'échelle ou d'unité et pourtant on va chercher à trouver l'impact de chacun d'entre eux sur notre concentration effective de gaz dans l'atmosphère.

On peut remarquer que les données NAEI représentent le même phénomène que celui que l'on cherche à qualifier et après plusieurs tests, on peut lever l'hypothèse que les deux données sont liées de façon linéaires étant donné qu'il s'agit de relevés de concentration dans l'atmosphère.

Ensuite, on cherche à qualifier l'impact des données spatiales, temporelles et d'altitude sur nos

concentrations de gaz. Ces facteurs sont donc ajoutés à l'équation en tant que paramètre de lissage car leur relation avec la concentration en gaz ne peut pas être décrite de façon linéaire.

Finalement, on cherche à voir l'effet combiné de plusieurs paramètres comme l'espace et le temps ou entre l'espace et l'altitude sur la concentration en gaz dans l'atmosphère. Nous faisons donc intervenir des tenseurs espace-temps et espace-altitude qui s'ajoutent au modèle.

L'équation finale pour un modèle gaussien si tous les paramètres disponibles sont inclus par l'utilisateur est la suivante :

$$\begin{aligned} concentration = \beta 0 + (NAEIdata)\beta 1 + f(longitude, latitude) + f(temps) + f(altitude) \\ + ti((longitude, lattitude), temps) + ti((longitude, lattitude), altitude) \\ + \epsilon, \epsilon \tilde{N}(0, \sigma 2) \end{aligned} \quad (3.5)$$

Grâce à la librairie mgcv disponible avec le langage R, la ligne de commande pour réaliser cette opération est la suivante :

```
Modèle_gam = mgcv::gam(formula = "pr_tc ~", formula_for_linear_data, " s(lon, lat)+
                                        s(ndate) +
                                         s(elevation)+
                                        ti(lon, lat, ndate, d=c(2,1)) +
                                         ti(lon,lat, elevation, d=c(2,1))" )),
                                          data = mydata)
```

A la sortie de cette opération, nous obtenons donc un modèle additif généralisé rassemblant toutes les interactions entre nos données et réutilisable afin de réaliser une interpolation sur grille comme décrit dans partie suivante.

3.2 Interpolation et exports

3.2.1 L'interpolation

Nous arrivons ici à la partie permettant de passer d'une acquisition sous forme de points répartis irrégulièrement sur l'Ecosse à une surface régulière dont on peut choisir la résolution spatiale.

Notre interpolation se fait sur une grille modulable à trois dimensions. La fonction expand.grid fournie par le package de base de R permet de créer une grille avec deux dimensions x et y représentant les valeurs de latitude et longitude et une dimension z représentant la variable de temps (cf. fig.3.4).

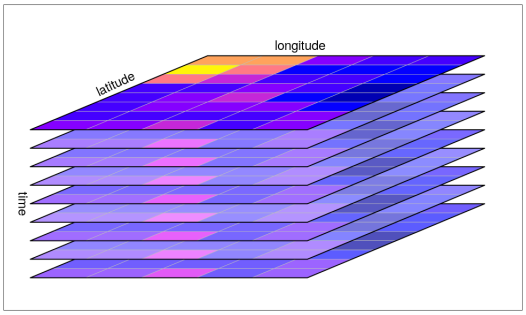


FIGURE 3.4 – Représentation de la grille d'interpolation

Voici un exemple d'utilisation de la fonction expand.grid en langage R :

```
interpolation_grid = expand.grid(
      lat= seq(min(mydata$lat),
               max(mydata$lat),
               by=0.02),
      lon = seq(min(mydata$lon),
                max(mydata$lon),
                by=0.02),
      ndate = seq(min(mydata$ndate),
                  max(mydata$ndate),by=(1))
```

Comme on le voit dans l'application de cette fonction, on peut choisir la résolution spatiale de la grille que l'on créé en séquençant en parties égales la latitude et la longitude. De plus, le séquençage de la dimension de temps permet de d'interpoler les données sur une grille dont les couches peuvent être des jours, semaines ou mois.

Une fois la grille créée avec les paramètres de résolution spatiale et de séquençage du temps choisis par l'utilisateur, on peut prédire la valeur de la concentration en gaz pour chaque nœud grâce au modèle statistique déterminé précédemment. Cette opération est réalisée grâce à la fonction predict.gam du package mgcv.

Nous obtenons donc en sortie une grille de valeurs de concentration pour un gaz donné pour chaque point dans les trois dimensions.

3.2.2 Présentation et exports

En sortie de cette phase d'analyse par modèle statistique et d'interpolation, les possibilités d'export et de présentation des résultats pour l'utilisateur sont multiples. Le plus important est bien évidemment la carte de concentration en gaz dont on peut étudier l'évolution en fonction des jours, semaines ou mois. La figure 3.5 ci-dessous montre un exemple d'export pour la concentration journalière en monoxyde de carbone pour le mois de mai 2021.

Ces cartes au-delà d'être tracées peuvent être exportées par un utilisateur au format .tif[2], un format d'export classique pour des données raster. L'utilisation potentielle de ces données raster est

2. Tagged Image File Format, type de format de partage de données raster

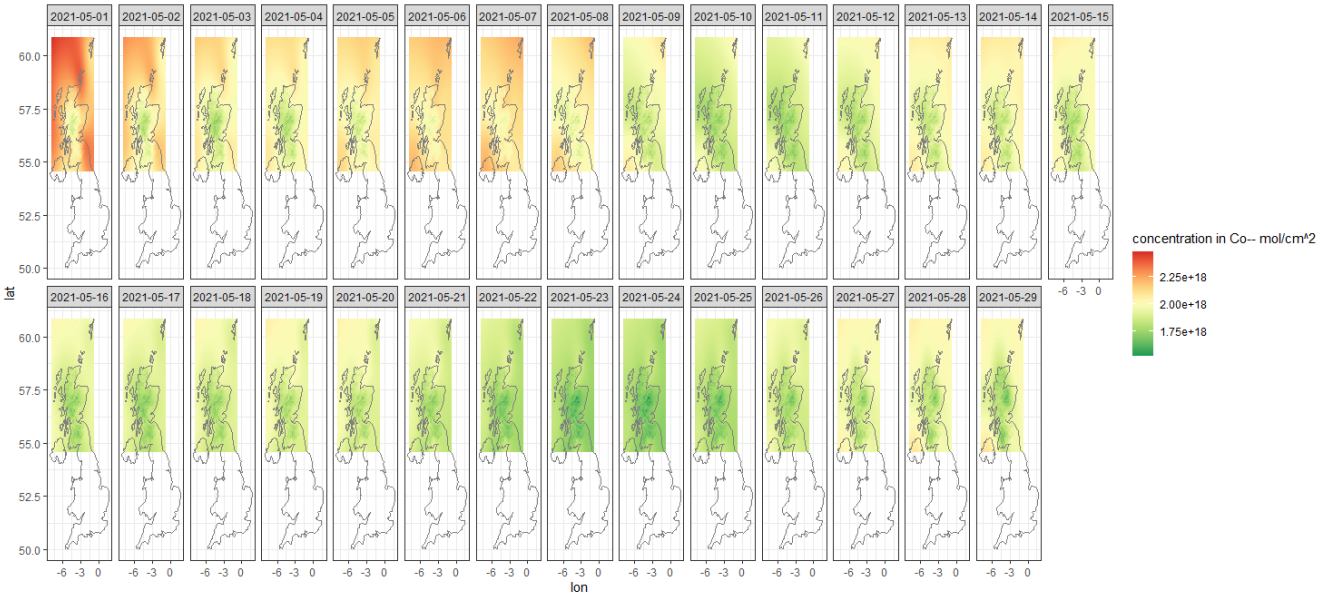


FIGURE 3.5 – Carte de concentration en CO pour le mois de mai 2021 - Ecosse

discutée dans la partie 3.3.2 de discussion et de continuation du projet.

Un autre type d'export possible pour l'utilisateur réside dans les résultats fournis par le calcul du modèle additif généralisé. En effet, le package mgcv permet de ressortir des données statistiques et de visualiser les interactions entre les différents paramètres. Ci-dessous, sur la figure 3.6 on retrouve un exemple du type d'export possible décrivant l'effet partiel du paramètre de temps pour le même jeu de données que pour la figure 3.5.

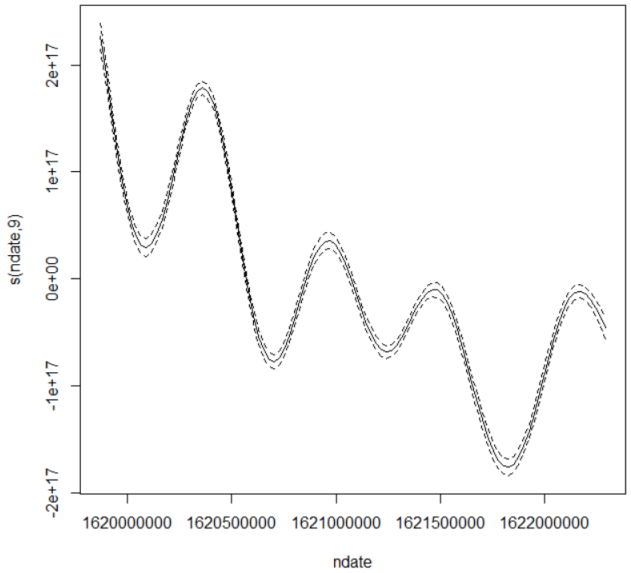


FIGURE 3.6 – Effet partiel sur le modèle du paramètre de temps. Produit = CO et dates = 2021/05/01 - 2021/05/29

Enfin, la librairie mgcv permet de visualiser sous forme de graphiques des relations allant jusquà trois dimensions comme ci-dessous (cf. 3.7) la relation entre les données spatiales longitude et latitude

et un paramètre linéaire.

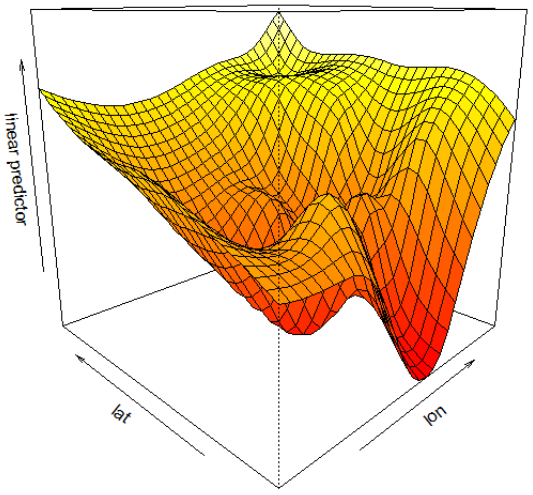


FIGURE 3.7 – Comparaison de l'effet des paramètres spatiaux sur le prédicateur linéaire. Produit = CO et dates = 2021/05/01 - 2021/05/29

3.3 Résultats et discussion

3.3.1 Résultats et retour d'expérience

Le résultat principal de ce stage est donc la mise à la disposition de l'Institut James Hutton d'un code complet permettant via des études statistiques et une validation des données de produire des cartes de concentration en gaz dans l'atmosphère au-dessus de l'Ecosse à partir de données SENTINEL 5p. La documentation du code qui permet de comprendre la parametrisation ainsi que les retours de chaque fonction pour un utilisateur est disponible en ANNEXE E.

En prenant du recul par rapport au projet, les travaux réalisés durant ces trois mois de stage vont participer à l'amélioration des capacités de compréhension de relations et de modèles complexes avec de très nombreuses variables que représentent les mesures et l'analyse des concentrations en gaz dans l'atmosphère. Seule une meilleure compréhension des phénomènes complexes liés aux émissions humaines et à leurs mouvements dans l'atmosphère peut amener à lutter efficacement contre le changement climatique notamment, mais aussi permettre de protéger les populations des effets néfastes sur leur santé des gaz polluants. L'inclusion de donnée satellite dans les modèles de surveillance de pollution des institutions à travers le monde est grandissante et ouvre de nouvelles possibilités notamment sur la surveillance à de très larges échelles.

D'un point de vue personnel, ce stage m'a permis de mettre en action mes capacités en géomatique dans un but écologique et humaniste. J'ai beaucoup appris sur les données environnementales mais aussi sur l'application des statistiques afin de représenter des modèles physiques. J'ai pu découvrir un nouvel outil, le langage R qui même, s'il n'est pas le plus adapté pour traiter de très larges jeux de données et des données géographiques, excelle dans ses capacités de calcul et son large choix

de fonctions pour les études statistiques.

3.3.2 Discussion et possibilités d'améliorations

Cette partie a pour but de cerner les points d'amélioration possibles du code et d'en prévoir de futures possibles utilisations.

Le premier point à aborder est l'ajout de plus de données issues de la mission SENTINEL 5p. En effet, par souci de temps, je ne me suis penché lors de ce stage que sur une fraction des données qu'il est possible d'extraire des fichiers Netcdf fournis par l'ESA.

Un point qui a été abandonné en cours de stage est notamment celui d'utiliser les noyaux moyens (averaging kernels en anglais). Ces noyaux moyens permettent de différencier selon l'altitude dans la colonne de gaz mesurée la probabilité de répartition des gaz. Pour être plus clair, les concentrations recueillies grâce au programme qui a été présenté dans ce rapport sont des données en deux dimensions. La concentration que l'on retire du HUB Copernicus représente la somme totale de tous les gaz (en mol/m^2) présents dans la colonne à la verticale de la zone étudiée. Il est donc impossible avec ce code de connaitre la concentration exacte en gaz au niveau du sol par exemple. Les noyaux moyens permettent d'obtenir une probabilité de retrouver une part de cette concentration totale pour chaque niveau d'altitude. Cela permettrait notamment une représentation en trois dimensions du déplacement des masses de gaz dans l'atmosphère.

D'autres données mises à la disposition de l'utilisateur par la mission SENTINEL 5p pourraient être utilisées selon les cas d'utilisation.

Dans un second temps, il faut préciser que l'utilisation des modèles additifs généralisé permet une grande modularité du code si l'on souhaite consolider l'interpolation en y ajoutant des jeux de données issus de sources différentes. On peut imaginer croiser avec le modèle déjà présent des données de températures, d'autres données de concentrations similaires à celles fournies par le NAEI ou bien d'autres. Un ajout intéressant serait celui de données anthropologiques pour réaliser dans quelle mesure la concentration des humains et des activités dans les villes par exemple a un effet prononcé sur la répartition des gaz polluants.

D'un point de vue d'ajout de données extérieures, le code reste donc très ouvert.

Un autre point à aborder est un des objectifs non remplis lors de ce stage qui était celui de mettre en place une application WEB à partir du code en utilisant le modèle R-shiny. Les raisons de l'abandon de cette piste ont déjà été abordées mais pour la suite, on peut imaginer que sur des ordinateurs plus puissants ou en précalculant la majorité des résultats intermédiaires chronophages, une application fluide puisse être mise en place. C'est une piste que Bethany WILKINS garde à l'esprit et sur laquelle elle va continuer de travailler après la fin de mon stage.

Finalement, dans la continuation du projet, une fois les rasters produits, la mise en place de statistiques pixel par pixel pourrait faire ressortir des tendances temporelles et d'observer l'évolution statistique des concentrations pour certains points spécifiques dans l'espace.

Conclusion

Afin de conclure ce rapport, nous pouvons rappeler les objectifs initiaux du stage. Il s'agit donc de participer à la réalisation pour le compte du l'institut James Hutton d'une application permettant de représenter des données environnementales appliquées à la surveillance des vulnérabilités des populations en Ecosse.

L'objectif principal était de coder des fonctions permettant de charger et de traiter, grâce aux librairies fournies par le langage R, des données issues de la mission SENTINEL 5p. A partir de ces données, nous devions retenir celles utiles dans un but de réaliser des cartes de concentration de gaz dans l'atmosphère.

Une fois l'obtention de données de concentration géolocalisées et temporellement renseignées, nous nous sommes concentrés sur l'ajout de données dans le but de calculer un modèle statistique permettant de décrire l'ensemble des interactions entre tous les paramètres.

Pour chaque point dans l'espace et le temps retiré des acquisitions satellite, nous avons assigné une valeur altimétrique mais aussi une valeur de concentration servant de référence issue des mesures effectuées par le NAEI sur l'année 2018.

Vient ensuite l'étude statistique par modèle additif généralisé. Ce modèle donne une information générale de répartition des valeurs et permet donc l'interpolation de la concentration en gaz polluants sur une grille en deux dimensions, à laquelle on rajoute une dimension temporelle.

Les résultats de cette analyse permettent donc d'extraire des rendus statistiques et graphiques. L'export le plus important restant bien évidemment les cartes de concentration par intervalle de temps comme le montre la Figure 3.5 en page 30.

On peut en conséquence considérer que les objectifs principaux du stage ont été remplis, même si la mise en forme, sous les traits d'une application WEB, n'a pas pu être réalisée. Des objectifs complémentaires peuvent encore être réalisés dans une phase de continuation potentielle du projet. Ces objectifs, comme l'ajout de plus de données au modèle statistique ou bien la réalisation d'une étude pixel par pixel, seront repris par Bethany WILKINS jusqu'à ce qu'elle arrive à un résultat complet permettant peut-être une diffusion grand public des travaux.

Bibliographie

Webographie / Bibliographie

[1] United States Environmental Protection Agency. *Climate Change Indicators : Atmospheric Concentrations of Greenhouse Gases*. https://www.epa.gov/climate-indicators/climate-change-indicators-atmospheric-concentrations-greenhouse-gases. Consulté en 2021.

[2] NOEL CRESSIE CHRISTOPHER K. WIKLE ANDREW ZAMMIT-MANGION. *Spatio-Temporal Statistics with R*. CRC PRESS, 2019.

[3] CLMS. *Page du site Copernicus permettant de télécharger des données altimétriques*. https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1. Consulté en 2021.

[4] ESA. *Page de la mission SENTINEL 5p regroupant les détails techniques*. https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-5p/products-algorithms. Consulté en 2021.

[5] James Hutton Institute. *About History Page*. https://www.hutton.ac.uk/about/history. Consulté en 2021.

[6] James Hutton Institute. *About ICS Page*. https://www.hutton.ac.uk/research/departments/information-and-computational-sciences. Consulté en 2021.

[7] James Hutton Institute. *Annual Review 2018-2019*. https://www.hutton.ac.uk/sites/default/files/files/publications/Hutton-Annual-Review-18-19.pdf. 2019.

[8] James Hutton Institute. *Annual Review 2019-2020*. https://www.hutton.ac.uk/sites/default/files/files/publications/James-Hutton-Institute-Annual-Review-2019-2020.pdf. 2020.

[9] Mitchell Lyons. *Article concernant l'application des GAMs dans une vision environnementale*. http://environmentalcomputing.net/intro-to-gams/. Consulté en 2021.

[10] NAEI. *Page du site officiel du National Atmospheric Emission Inventory pour le téléchargement de données*. https://naei.beis.gov.uk/data/map-uk-das. Consulté en 2021.

[11] R project. *Site du langage R*. https://www.r-project.org/. Consulté en 2021.

[12] Dr. Noam Ross. *Page Github de cours et de présentation des GAMs appliqués à R*. https://noamross.github.io/gams-in-r-course/. Consulté en 2021.

[13] J. Schwalb-Willmann. *Page Github de la librairie GetSpatialData*. https://github.com/16EAGLE/getSpatialData. Consulté en 2021.

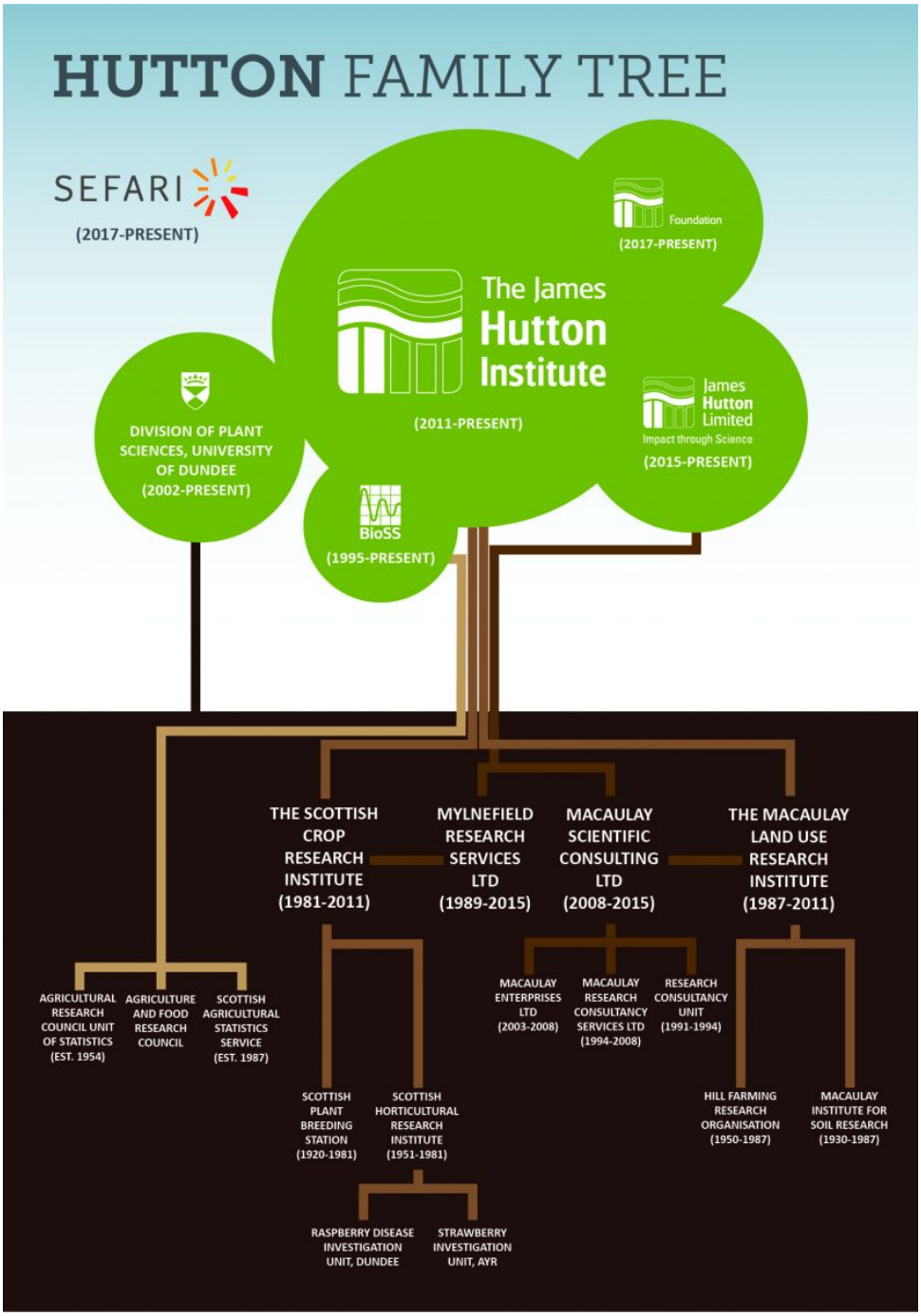
[14] Multiples utilisateurs. *Page d'aide et de documentation liée au programme SENTINEL*. https://docs.sentinel-hub.com/api/latest/data/sentinel-5p-l2/. Consulté en 2021.

Table des figures

Annexes

Arbre généalogique Institut James Hutton

Annexe A



Gestion

Annexe B

Figure B.1 – Diagramme de GANTT préliminaire du 02 juillet 2021

SUMMER INTERSHIP

Student : Charles Laverdure

ENVIRONMENTAL DEPRIVATION WEB APP USING SENTINEL 5P DATA

James Hutton Institute
Supervisor : Alessandro Gimona

Project Start: Mon, 5/24/2021
Display Week: 1

TASK	PROGRESS	START	END
Learning of the new tools			
R formation	100%	5/24/21	6/8/21
Sentinel 5P python package	100%	5/31/21	6/2/21
GetSpatialData R package	100%	5/31/21	6/4/21
Raster manipulation under R	100%	5/31/21	6/4/21
Bibliography and analitical study			
Copernicus product description	100%	6/8/21	6/28/21
Copernicus compressed data utilization	100%	6/8/21	6/15/21
Choice of useful datas for the project from the GetSpatialData results	100%	6/8/21	6/22/21
Test of non-statistical interpolations	100%	6/11/21	6/14/21
Test of statistical interpolations	100%	6/14/21	6/28/21
Cross validation to determine final use for interpolation	100%	6/28/21	7/1/21
Temporal interpolation	100%	6/30/21	7/4/21
Choice of data quality usage	100%	6/8/21	6/29/21
Research the GAMs mathematical definition	100%	7/13/21	7/25/21
GAMs pllied to R	100%	7/18/21	7/30/21
Adding the different functionalities (coding part)			
Getting the netCD files from copernicus	100%	6/8/21	6/13/21
Transformation of the netCD files into csv, lon lat concentration	100%	6/9/21	6/15/21
Get gam models from Netcdf models	100%	6/14/21	7/14/21
Use the predict function to interpolate data	100%	7/24/21	8/2/21
Creation of graphs to get complementary/more precise info on gases concentrations	100%	7/29/21	8/23/21
Potential Bugs resolutions and deployment / writting of the user manual/internship report			
Debbuging	100%	7/29/21	8/10/21
Deployment	100%	8/10/21	8/18/21
Instruction manual	100%	7/24/21	8/20/21
Internship report and oral presentation preparation	100%	8/4/21	8/20/21

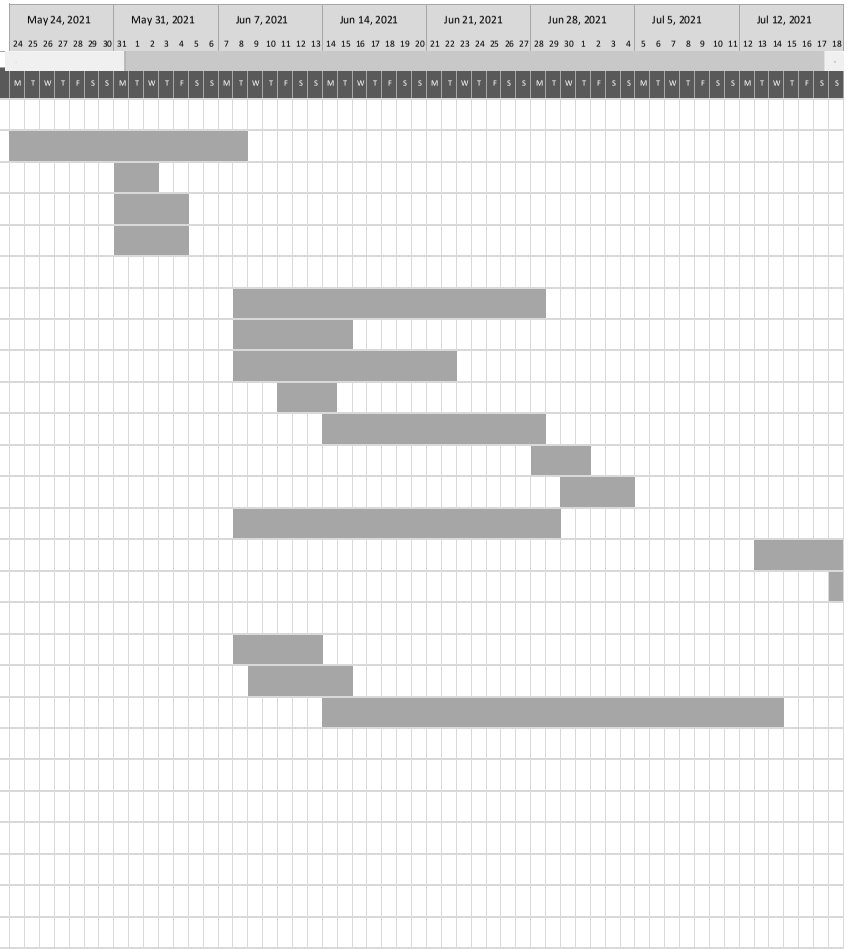


FIGURE B.2 – Diagramme de GANTT révisé du 11 août 2021

Liste des produits fournis par le satellite SENTINEL 5P

Annexe C

File type	Spectrometer	Spectral range [nm]	Comment
L1B_RA_BD1	UV	270 - 300	Radiance product band 1
L1B_RA_BD2		300 - 320	Radiance product band 2
L1B_RA_BD3	UVIS	320 - 405	Radiance product band 3
L1B_RA_BD4		405 - 500	Radiance product band 4
L1B_RA_BD5	NIR	675 - 725	Radiance product band 5
L1B_RA_BD6		725 - 775	Radiance product band 6
L1B_RA_BD7	SWIR	2305-2345	Radiance product band 7
L1B_RA_BD8		2345-2385	Radiance product band 8
L1B_IR_UVN	UVN	270-775	Irradiance product UVN module
L1B_IR_SIR	SWIR	2305-2385	Irradiance product SWIR module

Figure C.1 – Liste des produits disponibles pour le niveau 1 de la mission SENTNEL

Product type	Parameter
L2__O3___	Ozone (O_3) total column
L2__O3_TCL	Ozone (O_3) tropospheric column
L2__O3__PR	Ozone (O_3) profile
L2__NO2___	Nitrogen Dioxide (NO_2), total and tropospheric columns
L2__SO2___	Sulfur Dioxide (SO_2) total column
L2__CO____	Carbon Monoxide (CO) total column
L2__CH4___	Methane (CH_4) total column
L2__HCHO__	Formaldehyde (HCHO) total column
L2__CLOUD_	Cloud fraction, albedo, top pressure

FIGURE C.2 – Liste des produits disponibles pour le niveau 2 de la mission SENTNEL (1/2)

L2__AER_AI	UV Aerosol Index
L2__AER_LH	Aerosol Layer Height (mid-level pressure)
UV product[1]	Surface Irradiance/erythemal dose
L2__NP_BDx, x=3, 6, 7 [2]	Suomi-NPP VIIRS Clouds
AUX_CTMFC AUX_CTMANA	A-priori profile shapes for the NO2, HCHO and SO2 vertical column retrievals

FIGURE C.3 – Liste des produits disponibles pour le niveau 2 de la mission SENTNEL (2/2)

Liste des paramètres disponibles dans un fichier Netcdf pour le monoxide de carbon

Annexe D

	x
1	PRODUCT/delta_time
2	PRODUCT/time_utc
3	PRODUCT/qa_value
4	PRODUCT/latitude
5	PRODUCT/longitude
6	PRODUCT/carbonmonoxide_total_column
7	PRODUCT/carbonmonoxide_total_column_precision
8	GEOLOCATIONS/satellite_latitude
9	GEOLOCATIONS/satellite_longitude
10	GEOLOCATIONS/satellite_altitude
11	GEOLOCATIONS/satellite_orbit_phase
12	GEOLOCATIONS/solar_zenith_angle
13	GEOLOCATIONS/solar_azimuth_angle
14	GEOLOCATIONS/viewing_zenith_angle
15	GEOLOCATIONS/viewing_azimuth_angle
16	GEOLOCATIONS/latitude_bounds
17	GEOLOCATIONS/longitude_bounds
18	GEOLOCATIONS/geolocation_flags
19	DETAILED_RESULTS/processing_quality_flags
20	DETAILED_RESULTS/number_of_spectral_points_in_retrieval
21	DETAILED_RESULTS/pressure_levels
22	DETAILED_RESULTS/water_total_column
23	DETAILED_RESULTS/water_total_column_precision
24	DETAILED_RESULTS/semiheavy_water_total_column
25	DETAILED_RESULTS/semiheavy_water_total_column_precision
26	DETAILED_RESULTS/scattering_optical_thickness_SWIR
27	DETAILED_RESULTS/height_scattering_layer

Figure D.1 – (1/2)

28 DETAILED_RESULTS/surface_albedo_2325
29 DETAILED_RESULTS/surface_albedo_2335
30 DETAILED_RESULTS/wavelength_calibration_offset
31 DETAILED_RESULTS/chi_square
32 DETAILED_RESULTS/degrees_of_freedom
33 DETAILED_RESULTS/number_of_iterations
34 DETAILED_RESULTS/column_averaging_kernel
35 DETAILED_RESULTS/methane_total_column_prefit
36 DETAILED_RESULTS/methane_weak_twoband_total_column
37 DETAILED_RESULTS/methane_strong_twoband_total_column
38 DETAILED_RESULTS/water_weak_twoband_total_column
39 DETAILED_RESULTS/water_strong_twoband_total_column
40 INPUT_DATA/surface_altitude
41 INPUT_DATA/surface_altitude_precision
42 INPUT_DATA/surface_classification
43 INPUT_DATA/instrument_configuration_identifier
44 INPUT_DATA/instrument_configuration_version
45 INPUT_DATA/scaled_small_pixel_variance
46 INPUT_DATA/eastward_wind
47 INPUT_DATA/northward_wind
48 INPUT_DATA/surface_pressure
49 QA_STATISTICS/carbonmonoxide_total_column_histogram_axis
50 QA_STATISTICS/carbonmonoxide_total_column_pdf_axis
51 QA_STATISTICS/carbonmonoxide_total_column_histogram_bounds
52 QA_STATISTICS/carbonmonoxide_total_column_pdf_bounds
53 QA_STATISTICS/carbonmonoxide_total_column_histogram
54 QA_STATISTICS/carbonmonoxide_total_column_pdf

FIGURE D.2 – (2/2)

Documentation utilisateur du code

Annexe E

Developer documentation for environmental deprivation code

Version 1.0

Author Charles Laverdure Charles.laverdure02@gmail.com

Maintainer _____

Title Centralized and parametrizable code for the download and analysis of Sentinel 5p data over Scotland

Description This code allows a user to download sentinel 5p data from the sentinel hub and then apply different modification with the goal of interpolating the data over Scotland and extract statistics over a chosen period of time.

Depends R (>= 3.6.0), nlme (>= 3.1-64)

Imports base, ggplot2, mgcv, gstat, maps, dplyr, viridis, tidyverse, tiff, raster, sf, sp, ncdf4, stringr, tidyr, getSpatialData, data.table, stringi.

Date/Publication 08/10/21

List of files:

- main.R
- code_get_spatial_data.R
- conversion_to_csv.R
- add_data_to_dataframe.R
- Model_calculation.R
- plotting_and_statistics.R

List of associated functions:

- main.R → Main function that allows the centralized calling of all associated functions.

- code_get_spatial_data.R → download_sentinel_data()
- conversion_to_csv.R → convert_to_csv()
- add_data_to_dataframe.R → add_data_to_df()
- Model_calculation.R → model_calculation() and test_for_df_type()
- plotting_and_statistics.R →

(See graph for relation between functions (fig. below))

Description of functionalities per function:

Download_sentinel_data *Download Sentinel 5p data and transform to csv*

Description

Performs the downloading and saving of sentinel 5p data in the netcdf format over a given area and between specified dates. The user then chooses which variables of interest want to be kept as well as a minimum quality for each sub-area retrieved. The main goal of this function is to focus the retrieval and go from netcdf files that are 500Mo for 1/3 of a day to a more manageable csv.

Usage

dowload_sentinel_data (`product_wanted, vector_cordinates_aoi, date_of_ingestion, where_to_save_the_netcdf_files, quality_factor = 0.7, where_to_save_the_csv_files, copernicus_username`)

Type of return(s)

Dataframe and csv file

Arguments

`product_wanted` The product you want to retrieve available from the sentinel 5p hub. Documentation available at the address below for each product: https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-5p/products-algorithms

For the purpose of this code, only data with a total concentration column is available. The parameter has to be entered with the ESA nomination as a character chain. Choose from list below: `"L2__NO2___","L2__CH4___","L2__SO2___","L2__O3____", "L2__HCHO__", "L2__CO____"`.

`vector_cordinates_aoi` The coordinates between which the data should be retrieved. Entered as a simple matrix with the WGS 84 longitude and latitude extremities of the area as components: `c(min_latitude, max_latitude, min_longitude, max_longitude)`

Test were made with the coordinates `c(54, 61.2, -8.5, 0.5)` To cover the extent of Scotland.

`date_of_ingestion` Dates between which the data should be retrieved. See above ESA documentation to see the maximum range of dates available. Entered as a simple matrix of format `c("start_date", "end_date")`. Dates should be entered in the format YYYY/MM/DD.

`where_to_save_the_netcdf_files`

Path to where the user wants to save the temporary netcdf files on the computer. Should be written as a full path to a dummy folder where the files will be written and then erased as they are treated.

`Quality_factor` The quality factor is a parameter given inside the netcdf files provided by the Sentinel 5p hub. See documentation above for exact description on how to use this parameter. It is set automatically to 0.7 if no value is provided. Quality factor is comprised between 0 (no data) and 1 (perfectly retrieved data).

`where_to_save_the_csv_files`

Path to where the user wants to save the final csv files on the computer. Should be written as a full path to a folder where the files will be written day after day. As the code loops over the days, a new file is written including the new additions and the previous one is erased.

`copernicus_username` The username that the user has to enter to download the sentinel 5p data. A free account can be created at the following address:

https://scihub.copernicus.eu/dhus/#/self-registration

Details

This function works as a loop over the days. It is divided in two main parts: the downloading and opening of the necdf files and a part of treatment to convert the wanted data into a csv file.

See ESA product description for each product at the following address: https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-5p/products-algorithms

For the purpose of this code, offline data (OFFL) data is chosen to have pre-treated data with more available parameters. The user has to login the Copernicus Hub with it's own account. The username is to be entered on line 84 of the file code_get_spatial_data.R and then the password is required to continue the download.

Once a user has identified, the downloading will start. As a way to optimise the memory allocated to the calculations, the use can change the number of files downloaded between each phase of treatment. The parameter num_divisions represent the number of days the program will loop over between two treatment and thus the ereasement of the temporary netcdf files. 3 days is the standard value wich represents approximately 4.5 Go of data for each iteration. If the user has more space available, a larger num_division will slightly improve the calculation time.

As enunciated before, netcdf files from the Copernicus 5p program are voluminous and require a good internet connection to download a full set.

Temporary Netcdf files will be saved inside the dummy folder with the following denomination:

dummy_folder_name/Product_name___/sentinel-5p/official_copernicus_denomination.nc

Csv files as an output will be save in the output folder with the following denomination:

Csv_outputs/Product_name___/data_from_start_data_to_end_date.csv

Warnings

As mentioned before, the netcdf files are very voluminous so the user should have a good amount of free space available and an unlimited wifi connection. In case of a bug, verify that the product name are written according to the specified format indicated above.

Examples

```
library(getSpatialData)

library(ncdf4)

library(data.table)

library(stringi)

library(tidyverse)

library(sf)


dowload_sentinel_data (product_wanted = "CO____",

                  vector_cordinates_aoi   =   c(54,   61.2,   -8.5,   0.5),
                  date_of_ingestion   =   c("2020-06-02",   "2020-06-07"),
                  where_to_save_the_netcdf_files=
                  "C:/Users/random/Documents/netcdf_outputs",
                  quality_factor = 0.7,

                  where_to_save_the_csv_files=
                  "C:/Users/random/Documents/csv_outputs",

                  copernicus_username = "random_user123")
```

References

[1] https://notes.stefanomattia.net/2018/02/14/Plotting-Sentinel-5P-NetCDF-products-with-R-and-ggplot2/

[2] https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-5/data-products

[3] https://forum.sentinel-hub.com/

[4] https://atmospherictoolbox.org/

convert_to_csv *Transform data from netcdf format to csv*

Description

This function is associated to the above get_spatial_data function. It allows for the opening of the previously downloaded netcdf files, chooses which parameters to keep and saves the aggregated dataframe as a csv file.

Usage

convert_to_csv (`quality_factor, vector_cordinates_aoi, directory, where_to_save_the_csv_files, num_prod, list_total_files, df_to_return`)

Type of return(s)

Dataframe, number for product and list of treated files.

Arguments

`quality_factor`	The quality factor is a parameter given inside the netcdf files provided by the Sentinel 5p hub. See documentation above for exact description on how to use this parameter. It is set automatically to 0.7 if no value is provided. Quality factor is comprised between 0 (no data) and 1 (perfectly retrieved data)
`vector_cordinates_aoi`	The coordinates between which the data should be retrieved. Entered as a simple matrix with the WGS 84 longitude and latitude extremities of the area as components: `c(min_latitude, max_latitude, min_longitude, max_longitude)` Test were made with the coordinates `c(54, 61.2, -8.5, 0.5)` To cover the extent of Scotland.
`directory`	The directory to the netcdf files location. It is given automatically when created in the get_spatial_data function.
`where_to_save_the_csv_files`	Path to where the user wants to save the final csv files on the computer. Should be written as a full path to a folder where the files will be written day after day. As the code loops over the days, a new file is written including the new additions and the previous one is erased.
`Num_prod`	As the naming differs for each products, the Num_prod parameter allows the user to choose a product from the netcdf file summary. This number allows the user to choose a product (Total column is recommended). See the product description and documentation above for more information on available parameters for each products. A possible improvement of the code would be for the user to be able to

	choose multiple products to add to the dataframe returned from the function.
`list_total_files`	This variable is here for redundancy, it is used so that two similar netcdf files are not used two times. It returns for each call of the conversion_to_csv function the list of the names of the files already treated to the get_spatial_data function. This is made because the erasing of the files post treatment doesn't work at every loop.
`df_to_return`	This is the dataframe that is building up between every loop as the

Details

This function fully includes itself into the get_spatial_data() function. It is used in a loop to open and treat the netcdf files.

A full list of retrievable parameters is given as an example in annex and see the figure 1.1 below to understand the composition of Copernicus sentinel 5p netcd files.

This function allows for a user choice of a parameter to add to the returned dataframe. At this stage in the code, the main parameter of interest is the total concentration column but in a further version, the choice could be given to the user to add more parameters provided by each netcdf files. Other than the total concentration column, the program automatically retrieves the values for longitude, latitude, the UTC time for each individual acquisition and the qa value, explained above.

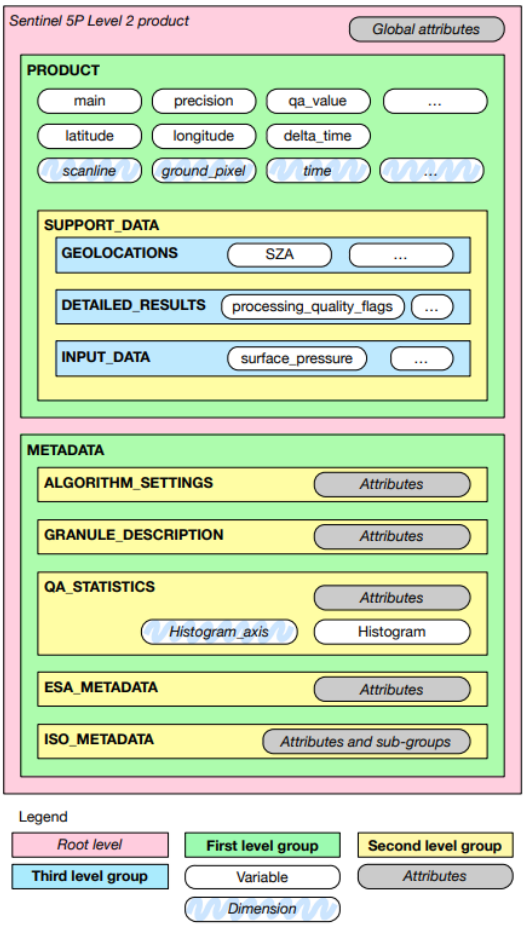


Figure 1.1 : Graphical description of the generic structure of a Level 2 file. The elements labelled as a dimension are coordinate variables. (Sentinel 5p product user manual)

Warnings

No particular warning should araise from this function if none is risen for the get_sentinel_data(). The user should be careful in entering it's product number when asked by the program.

Examples

```
library(ncdf4)

library(stringi)

library(tidyverse)


convert_to_csv(quality_factor = 0.7,

               vector_cordinates_aoi = c(54, 61.2, -8.5, 0.5),

               directory = "C:/Users/random/Documents/netcdf_outputs",

               where_to_save_the_csv_files=
               "C:/Users/random/Documents/csv_outputs",

               num_prod = NULL,

               list_total_files = list(),

               df_to_return = dataframe())
```

References

[1] https://notes.stefanomattia.net/2018/02/14/Plotting-Sentinel-5P-NetCDF-products-with-R-and-ggplot2/

[2] https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-5/data-products

[3] https://forum.sentinel-hub.com/

[4] https://atmospherictoolbox.org/

[5] https://cran.r-project.org/web/packages/ncdf4/index.html

add_data_to_df *Add external data to a dataframe containing Sentinel 5p concentrations*

Description

This function allows the user to choose which data should be incorporated into the dataframe containing the sentinel 5p concentrations. Data available are elevation data from the European Copernicus DEMs reprojected into WGS84 to match our data. The other type of data is a median for gases concentrations established per emitter by the National Athmospheric Emission Inventory (NAEI) cropped for Scotland for 2018. Those two sets of data are available in the TIFF format (raster) in the folder "data" accompanying the rest of the code. (see parts below for name and description of those files). The final goal of adding those type of data will be clearer when calculating the GAM model in the next function. In fact, adding more data to a dataset yields to more interesting and overall accurate results if the datasets are correlated.

Usage

add_data_to_df (`product, df, elevation, full, naei, path_to_DEM, path_to_naei_tiff`)

Type of return(s)

Dataframe.

Arguments

`product`	The product that your dataframe represents the concentration of. Should be written as a character chain following the Copernicus sentinel 5p naming norms (ie. carbon monoxide = "CO____").
`df`	The actual dataframe containing the data from sentinel 5p netcdf files and extracted through the function get_spatial_data() (see above).
`elevation`	Boolean, set to 1 if you want to add elevation data, 0 if you don't.
`full`	Boolean, set to 1 if you want to add elevation data from all 3 files available in the "data" folder, 0 if you want data from just one DEM. (explanation available in the details part below).
`naei`	Boolean, set to 1 if you want to add naei yearly data (only available for CO, SO2 and CH4, set to 0 for all other products), 0 if you don't want to add this data.
`path_to_DEM`	Path to the DEM tiff files as a list (c(DEM1, DEM2,…)). Should be written as a full path to make sure no bug arises.
`path_to_naei_tiff`	Path to the naei tiff folder, no need to precise which one, the right file will be found using the product name. Should be written as a full path to make sure no bug arises.

Details

This function is dependant on external raster files in the TIFF format. Data available for elevation are eu30_DEM_WGS.tif, eu40_DEM_WGS and Scotland_DEM_WGS. The first two have followed the same transformations and represent the south and north part of Scotland in two parts. All three raster files are downloaded from the Copernicus hub at the following address: https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1. The changes applied to crop and put the data into the right CRS are available in the code prepare_external_data.R.

We can observe small differences post-treatment between the two sets of DEMs : Scotland_DEM_WGS and the two other, that is why we've let the choice to the user to add data from just one or from all the DEMs before making a final average into one column.

Other than elevation data, the choice is given to the user to use data from the National Athmospheric Emission Inventory (NAEI) given for the year 2018. The NAEI only provides data for three types of pollutants that match the one we can extract in the get_spatial_data.R file: CO, SO2 and CH4. All the relative info and where to download the original ASCII files can be found at the address below: https://naei.beis.gov.uk/data/map-uk-das

In the code, space is let to the user in the comments to potentially use the "terra" package in R which is an augmented version of the "raster" package but is unstable.

Warnings

The warnings here concern the computation time, the raster::extract function in R can be very heavy to run, so for an optimized computation time, it is recommended to set the parameter "full" to 0 to reduce by 2 the number of time this function is used.

Examples

```
library(dplyr)

library(tiff)

library(raster)

library(sf)

#library(terra)

add_data_to_df(product = "CO____",

                df = dataframe(),

                elevation = 1,

                full = 0,

                naei = 1,

                path_to_DEM      =      c(C:/some_dir/Scotland_DEM_WGS.tif,
C:/some_dir/eu30_DEM_WGS.tif,
C:/some_dir/eu40_DEM_WGS.tif),

                path_to_naei_tiff = C:/some_dir/folder_for_naei_data)
```

References

[1] https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1

[2] https://naei.beis.gov.uk/data/map-uk-das

[3] https://cran.r-project.org/web/packages/raster/index.html

[4] https://cran.r-project.org/web/packages/raster/index.html

[5] https://cran.r-project.org/web/packages/tiff/index.html

[6] https://cran.r-project.org/web/packages/terra/index.html

---

model_calculation *Compute a GAM model for later interpolation and analysis*

Description

This function is the core of the whole program. It takes advantage of the R mgcv library to compute a GAM (General Additive Modeling) model. The GAM model allows to compute the gases concentration over Scotland in our case taking into account a large number of factors. Once this model is computed, it can be use further along the program to interpolate values for concentration but also to extract statistical relations between factors.

Usage

model_calculation (`dataf, time_period, minimum_na_prct`)

Type of return(s)

Dataframe, large gam model.

Arguments

`dataf`	The dataframe including the values of total column concentration with all the data added through the add_data_to_df() function.
`time_period`	The time period over which you wish the model to be calculate upon.
`minimum_na_prct`	Added value from the add_data_to_df() don't always cover the whole extent of our initial data. Thus, some areas are left with NA values and the model calculation can be complicated by this lack of data. This argument lets the user set a minimum amount of data that a specific column has to be accounted for in the GAM calculation. It

ranges from 0 to 1 with 0 meaning all the point should have a value and all NA values should be remove and 1 meaning that all NA values are inserted into the GAM calculation. After testing and documentation, the maximum value for this parameter should be 0.6.

Details

GAMs (General Additive Modelling) allow for the combination and computation of complex non-linear relationships between data. In our case, it is very interesting first because we talk about spatio-temporal data with non-linear relations. In a second hand, it also allows for the adding of other types of data on a different scale.

GAMs allow to fit model to complex set of data and then plot the results to see the relevance of each set of data on the total concentration per product. In our case, once the model is calculated, we store it with the goal of using it later on paired with a "predict.gam" function to interpolate values over a grid that will allow us to make comparisons on similar scales between the concentration values at certain places in time.

In general, a GAM formula in R is of the type gam(formula = parameter_of_interest ~ factor_1 + factor_2, data = mydata).

Factor_1 and factor_2 can be either included as linear in their relation to the parameter of interest or as a complex relationship or as a tensor interaction.

In our case, after testing and documentation, we've decided to present a GAM formula as follows:

Formula = (pr_tc ~ naei_data + s(time, space and elevation data) + ti(space against time and elevation data)

Pr_tc is the total_column concentration for the chosen product.

NAEI data are inherently linear when compared to pr_tc and represented as such.

Time, space and elevation are computed as complex relations with pr_tc (s())

Finally, time and elevation are computed as tensor interactions with longitude and latitude.

See [3] and [9] in references for methodology and more information about the techniques that yields to those types of formula.

Warnings

As for most of the code, this part of calculating the GAM model can be time consuming for large datasets. If nothing is returned, check the composition of you dataframe as this function is made to theoretically work for all combination of data resulting from the previous functions present in the code.

Examples

```
library(mgcv)

library(gstat)
```

```
library(dplyr)

library(tydiverse)

model_calculation(dataf = dataframe(),

                        time_period      =      c("2020-01-20","2020-03-29"),
                        minimum_na_prct = 0.6)
```

References

[1] https://cran.r-project.org/web/packages/mgcv/index.html

[2] https://spacetimewithr.org/Spatio-Temporal%20Statistics%20with%20R.pdf

[3]https://stats.stackexchange.com/questions/244042/trend-in-irregular-time-series-data/306361#306361

[4] https://m-clark.github.io/generalized-additive-models/

[5]https://stats.stackexchange.com/questions/35510/why-does-including-latitude-and-longitude-in-a-gam-account-for-spatial-autocorre

[6] http://environmentalcomputing.net/intro-to-gams/

[7] https://www.r-bloggers.com/2020/02/spatial-predictions-with-gams-and-rasters/

[8] https://noamross.github.io/gams-in-r-course/

[9]https://github.com/eric-pedersen/mgcv-esa-workshop/blob/master/example-spatio-temporal%20data.Rmd

plotting	*Plot the data extracted from the GAM model over Scotland*