



HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

HEIG-VD

TRAVAIL DE BACHELOR

La Terre de nuit vue de l'espace

Antoine FRIANT
Haute École d'Ingénierie et de Gestion du
Canton de Vaud
Yverdon-les-Bains, VD, CH
antoine.friant@gmail.com

26 juillet 2018

Table des matières

1 Cahier des charges	iii
1.1 Résumé du problème	iii
1.2 Objectifs	iii
1.3 Limitations	iv
1.4 Description fonctionnelle	iv
1.5 Délais	iv
2 Résumé	v
3 Introduction	1
4 Gleam	2
4.1 Installation	2
4.2 Notebooks	2
4.2.1 country_stats	2
4.2.2 gleam	3
4.2.3 rastercomparator	3
4.2.4 scraper	3
4.2.5 viirs_extractor	3
5 Exploration des données	4
5.1 Jeux de données	4
5.1.1 Images satellites	4
5.1.2 Grilles de population	7

5.1.3	Statistiques nationales	8
5.2	Recherche de corrélation	9
5.2.1	Statistiques nationales	9
5.2.2	Grille de population	14
6	Modèle	18
6.1	Environnement de développement	18
6.2	Réseau de neurones	18
6.3	Première topologie	19
6.4	Premiers résultats	19
6.5	Améliorations	21
6.5.1	Prétraitement	21
6.5.2	Topologie et paramètres	22
6.5.3	Post-traitement	22
6.6	Réseau définitif	23
6.7	Résultats	24
6.7.1	Validation croisée	24
6.7.2	Évaluation des prédictions	24
6.8	Discussion	24
6.9	Prédictions non validées	28
7	Conclusion	31
7.1	Discussion finale	31
7.2	Pour aller plus loin	31
8	Authentification	35

1 Cahier des charges

1.1 Résumé du problème

Les données géographiques sont nécessaires pour la prise de décisions importantes. Cependant la fiabilité et la disponibilité de ces données ne sont pas homogènes dans le temps et selon le lieu. Certaines de ces données ont une forte corrélation avec la lumière perçue par les satellites pendant la nuit.

Grâce à l'apprentissage automatique (*machine learning*), il est possible d'entraîner un réseau de neurones sur des données d'une date et d'un lieu connus pour reconstituer une carte de données géographiques à partir d'une image satellite nocturne.

Le travail à effectuer consiste à explorer différents types de données géographiques afin d'en choisir un, et faire de la prédiction sur ce type de données grâce à un réseau de neurones.

1.2 Objectifs

Le TB consiste dans un premier temps à explorer les données suivantes :

- Images satellites nocturnes de la Terre,
- Population humaine,
- Population animale,
- Densité végétale,
- PIB,

Et toutes autres données jugées pertinentes dans le but d'entraîner un réseau de neurones capable de prédire une estimation d'une donnée utile, à partir d'une image satellite de la terre de nuit.

La réalisation d'une application qui entraîne et exploite ce réseau de neurones est l'objectif de la seconde partie du TB.

Le but final est de pouvoir estimer, grâce au machine learning, des informations dont on ne possède pas de données à jour. Et cela à partir d'images satellites de nuit récentes, ou d'une combinaison de ces images avec une autre donnée à jour.

1.3 Limitations

L'application sera compatible avec Windows 10 et Archlinux, et nécessitera l'installation de librairies tierces (telles que Keras et TensorFlow). Elle ne possèdera pas nécessairement d'interface utilisateur.

L'utilisateur sera responsable de fournir les données à l'application dans un format supporté.

1.4 Description fonctionnelle

L'application prend en argument au moins deux jeux de données géographiques de format imposé : une image satellite nocturne et un autre type de données à déterminer au cours du projet. Après un long temps d'entraînement (une semaine au maximum, dépend de la machine utilisée), un modèle est généré.

Une fois le modèle généré, il est sauvegardé et réutilisable sur une autre image satellite nocturne (d'une date et/ou d'une région différente). Lorsque le modèle est appliqué sur une image satellite, une carte est recréée, affichant le résultat des prédictions.

Par exemple, si au cours du travail de bachelor il s'avère que la population par kilomètre carré est une donnée utile et utilisable, l'application devra prendre en argument une image satellite nocturne ainsi qu'une carte des populations de même taille et de même résolution pour entraîner le réseau de neurones. Une fois le modèle généré, l'application devra être capable de regénérer une approximation de la carte de population par kilomètre carré à partir d'une image satellite.

1.5 Délais

15 juin 2018 : Rapport intermédiaire

27 juillet 2018 : Rapport final et application fonctionnelle

Entre le 3 et le 14 septembre 2018 : Soutenance du travail de bachelor

2 Résumé

Les techniques d'apprentissage automatique offrent de nouvelles opportunités en matière de traitement de données volumineuses. Ce projet s'intéresse en particulier aux images satellites de la Terre de nuit. La disponibilité d'observations récentes en fait une excellente base pour les prédictions d'une intelligence artificielle.

Ce travail consiste à explorer ces images, ainsi que d'autres données géographiques avec lesquelles il serait possible de tirer un parallèle : la démographie, l'écologie, l'économie, etc. Une fois un type d'information choisi, il s'agit d'entraîner un réseau de neurones capable de générer une prédiction pour ces valeurs à partir d'une image satellite nocturne.

Le modèle ainsi entraîné est capable de prédire, par exemple, la répartition de la population en Colombie avec une résolution au quart de kilomètre carré. Cette alternative aux comptages, bien que moins fiable, est disponible dès lors que l'image satellite est publiée, car la génération d'une prédiction ne prend que quelques secondes. Celle-ci sera également de résolution plus élevée et bien moins coûteuse à produire.

3 Introduction

Les produits d'imagerie satellite sont devenus abondants et largement accessibles au cours des dernières décennies. De nombreux satellites prennent des photographies de la Terre chaque heure du jour *et de la nuit*. Les observations nocturnes révèlent des caractéristiques peu évidentes de jour, parfois même cachées. Les routes apparaissent, les villes montrent leurs lumières, même les bateaux de pêche aveuglent les océans avec des projecteurs pour attirer les poissons.

La disponibilité, la résolution et l'uniformité de la qualité de ces données contrastent fortement avec le manque de fiabilité d'autres informations géographiques utiles lors de prises de décisions importantes. Par exemple, la répartition de la population est une estimation précise en Suisse, mais très approximative au Kenya. D'autres mesures intéressantes incluent : la consommation en électricité, les émissions de C0₂, la couverture végétale et la présence de faune. Les lumières nocturnes observées depuis l'espace donnent des indications sur chacune de ces mesures alors qu'elles peuvent manquer dans une région à une date donnée.

Le but de ce projet est d'extraire autant d'informations que possible de l'imagerie satellite nocturne en utilisant l'apprentissage automatique (*machine learning*) sous la forme de réseau de neurones.

4 Gleam

Le projet Gleam ("lueur") se présente comme un ensemble de notebooks Jupyter disponible sur GitHub à l'adresse <https://github.com/Bertral/Gleam>. Il contient tous les scripts écrits au cours de ce travail. Les données brutes telles que les grilles de population, images satellites, cartes des frontières et statistiques officielles ne sont pas incluses dans le projet en ligne. Il est donc de la responsabilité de l'utilisateur de les obtenir depuis leur source d'origine, ainsi que de les traiter de manière à les rendre utilisables à l'aide de logiciels GIS (tels que QGIS) et des scripts du projet Gleam.

4.1 Installation

Les notebooks nécessitent la version 3.7 de Python. Ils ont été testés sur Windows 10 et Archlinux, mais devraient fonctionner sur toutes les versions de Windows et Linux.

Les librairies tierces `shapely`, `gdal`, `fiona` et `rasterio` sont compilées à partir de code C dans le but d'optimiser leur performance, mais cela a pour conséquence qu'elles ne sont compatibles qu'avec Linux. Les utilisateurs Windows doivent donc les télécharger et les installer manuellement (`pip install nomdupackage.whl`) depuis le site de Christoph Gohlke [10], qui regroupe un grand nombre de librairies scientifiques recompilées pour Windows. Ce sont les versions "cp37" qui sont compatibles avec Python 3.7.

Pour installer le reste des prérequis sur Linux et Windows, il suffit de lancer la ligne de commande `pip install -r requirements.txt` depuis la racine du projet.

Pour naviguer à travers les notebook sur Windows, il faut lancer la commande `python -m notebook`, ou `jupyter notebook` sur Linux.

4.2 Notebooks

Chaque notebook se trouve dans le répertoire portant son nom, à la racine du projet. Les fichiers générés par ces notebooks (cartes, graphes, logs et modèles) seront enregistrés dans le dossier du notebook.

4.2.1 country_stats

Ce notebook permet de compiler des statistiques sur la luminosité émise par chaque pays à partir d'une image satellite. Il produit ensuite une série de graphes montrant l'évolution de la

lumière vue de l'espace au fil des années pour chaque pays, ainsi qu'un nuage de point plaçant les pays selon leur lumière émise et leur population, PIB ou énergie consommée, colorés par développement économique.

Cette visualisation a pour but de chercher une corrélation à exploiter lors de la création du modèle.

4.2.2 gleam

Ce notebook est l'aboutissement du projet. Il entraîne un réseau de neurones à convolutions puis fait une prédiction de la répartition de la population sur une zone géographique à partir d'une image satellite nocturne.

4.2.3 rastercomparator

Ce notebook regroupe différentes manières de comparer deux fichiers GeoTIFF de même emprise : carte des différences par pixel, coefficient de corrélation et différence entre les sommes de pixels. Il est utile pour évaluer la qualité d'une prédiction ou observer l'évolution d'une carte d'une année à l'autre.

4.2.4 scraper

Ce script télécharge des images satellites fragmentées au format PNG et les assemble pour recréer des images globales. Comme le format GeoTIFF est largement plus adapté que le format PNG pour le travail à effectuer, ce notebook n'a pas servi pour le reste du projet.

4.2.5 viirs_extractor

Ce script répond au besoin spécifique d'extraire les archives volumineuses de la NOAA [7] contenant des images satellites haute résolution. Puis, il calcule une médiane annuelle pour chaque pixel des images mensuelles afin d'en éliminer les valeurs atypiques dues à l'absence occasionnelle d'observations sans nuages ou bruit.

5 Exploration des données

5.1 Jeux de données

5.1.1 Images satellites

NASA Worldview

La première source de données explorée est l'application "Worldview" de la NASA [14]. Elle permet de visionner un grand nombre d'images satellite composites sur un globe en trois dimensions (voir figure 5.1). Parmi les jeux de données disponibles sont trois jeux d'images nocturnes.

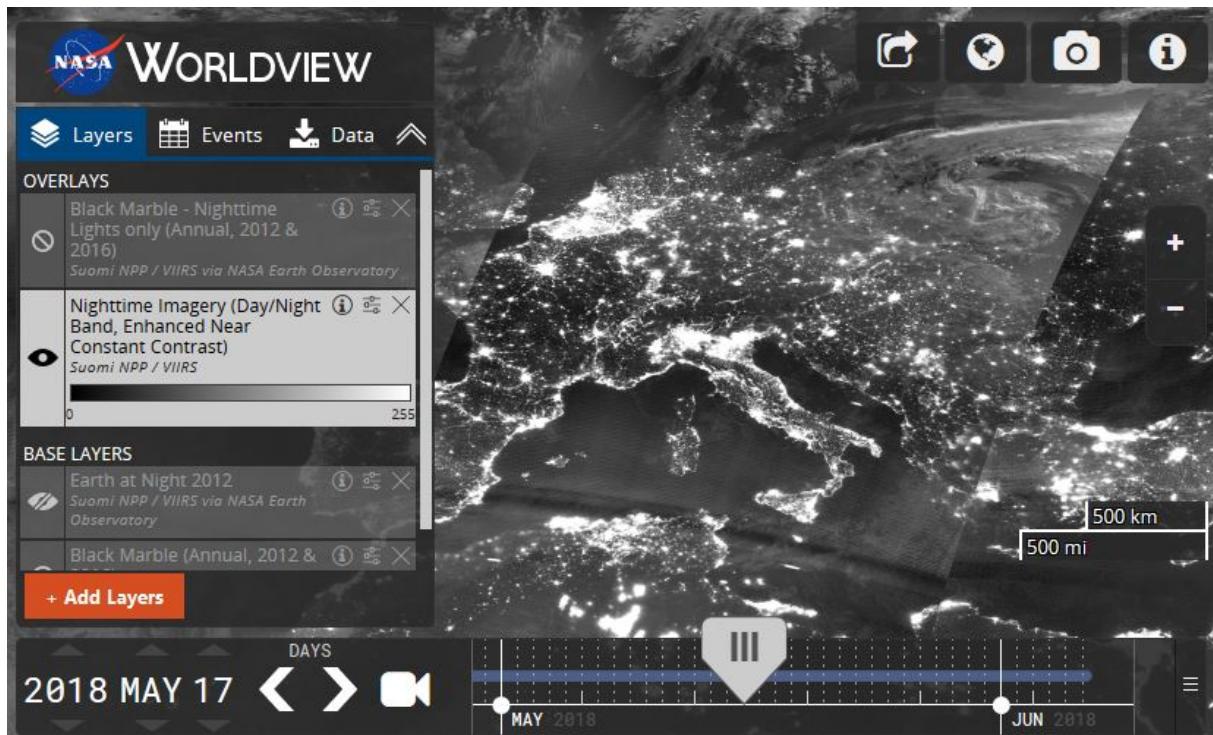


FIGURE 5.1 – Outil de visualisation NASA Worldview [14].

Le premier jeu de données est une série d'images composites capturées par le satellite Suomi NPP opéré par la NASA, la NOAA et le Département de la Défense des États-Unis. Il est mis à jour toutes les quelques heures, et présente une image composite chaque jour depuis le 30 novembre 2016. Elle possède deux défauts éliminatoires : la période d'observation actuellement disponible (à peine plus d'une année) n'est pas suffisamment longue pour observer une évolution

significative des villes depuis l'espace, et les images ne sont pas traitées. Cela signifie que celles-ci sont très fortement bruitées par les nuages et la lumière ambiante due aux différentes phases de la Lune.

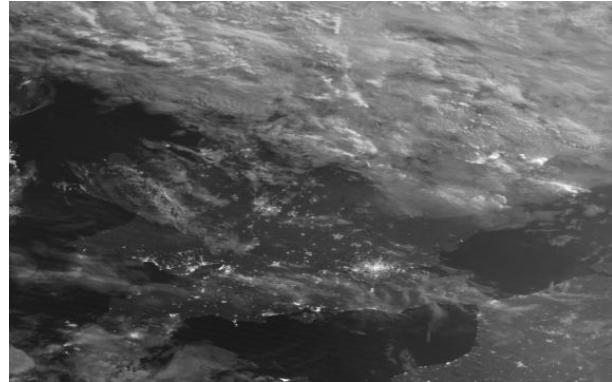


FIGURE 5.2 – Image satellite quotidienne servie par NASA Worldview [14], représentant la Grande-Bretagne et son climat nuageux.

Les deux autres jeux de données nocturnes sont des images composites : des clichés pris tout au long de l'année ont permis de fabriquer une seule image du globe dont la luminosité ambiante est constante (moyennée) et sur laquelle les nuages n'apparaissent pas. Malheureusement, l'outil Worldview ne permet pas un téléchargement direct de ces images *dans leur pleine résolution*. Heureusement, la NASA a mis à disposition une API REST (<https://wiki.earthdata.nasa.gov/display/GIBS/GIBS+API+for+Developers>) pour télécharger des "tuiles" de n'importe laquelle de leur image. Seulement le format PNG est disponible. Ce format ne contient pas d'informations géographiques, ce qui complique leur utilisation pour la suite de ce travail. Un script Python suffit pour télécharger et assembler les tuiles (figure 5.3) et reconstituer une image complète du globe de plus de 800 millions de pixels (figure 5.4).

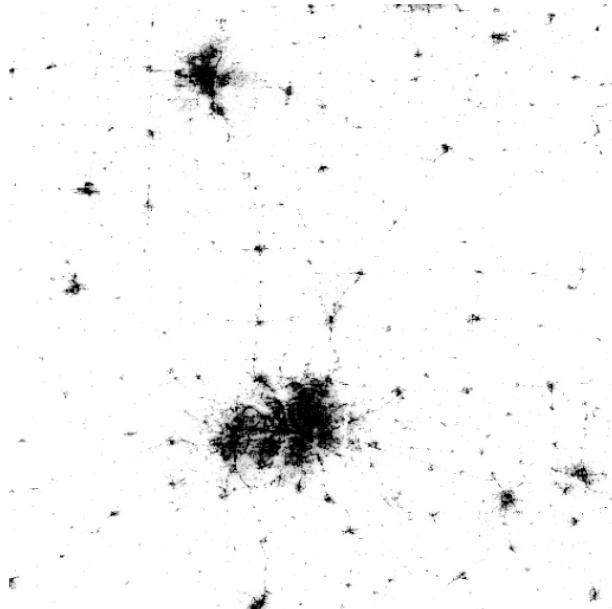


FIGURE 5.3 – Une tuile de l'image de 2016 montrant la ville de Dallas (USA) après avoir été mise en couleurs négatives.

Le script Python utilisé se trouve dans le notebook `scraper` et nécessite l'installation de la librairie Pillow pour le traitement des images, ainsi qu'urllib pour le téléchargement en soi. Son



FIGURE 5.4 – Image globale annuelle (2016) reconstituée à partir de tuiles téléchargées, puis mise en couleurs négatives.

exécution peut demander plus d'une heure pour le téléchargement (vitesse limitée par le serveur), et plus de 4 Go de RAM pour l'assemblage des tuiles.

Agence américaine d'observation océanique et atmosphérique

La source d'images satellites retenue pour la suite du travail est celle de l'Agence américaine d'observation océanique et atmosphérique (abrégé NOAA). Des images satellite nocturnes composites ("Average Lights X Pct") sont disponibles pour les années 1992 à 2013 [8] (1km par pixel, unité arbitraire de 0 à 63) et 2012 à 2018 [7] (mensuelles, 500m par pixel, radiance en nanowatts/cm²/sr), une période suffisante pour observer des changements depuis l'espace. De plus, ces données sont disponibles en format GeoTIFF, qui contient les informations géographiques nécessaires pour superposer cette carte sur une autre. Il est donc possible d'explorer et manipuler ces cartes à l'aide de logiciels libres tels que QGIS. Il s'agit en réalité des mêmes clichés fournis par Worldview (et Google Earth), mais plus nombreux et dans un format bien plus exploitable.

Ces images ont été créées en moyennant la valeur de luminosité de chaque pixel sur une année, en ignorant les pixels couverts par des nuages, et en multipliant cette moyenne par la fréquence de détection de lumière sur le pixel au cours de l'année. Les images mensuelles de 2012 à 2018, en revanche, possèdent des pixels pour lesquels il n'y a pas d'observation sans nuages ou lumière parasite. Pour éliminer ces valeurs atypiques, le notebook `viirs_extractor` créé une carte compilant la médiane annuelle de chaque pixel.

5.1.2 Grilles de population

Le Socioeconomic Data and Applications Center (SEDAC) [6] met à disposition des grilles de populations pour le monde entier, sous forme de fichier GeoTIFF. Chaque "case" de 1 km² est représentée comme un pixel et contient une estimation du nombre de personnes vivant dans cette case. QGIS s'est à nouveau montré d'une grande aide pour visualiser et manipuler ces données volumineuses (exemple en figure 5.5).

On remarque que la valeur de densité de population ne varie souvent pas à l'intérieur d'une sous-région. En effet, la qualité des données fournies par ces grilles est très variable selon les pays. Parmi les régions les plus détaillées, on trouve les USA, l'Italie, le Portugal et le Brésil. En plus de cette disparité, il faut garder à l'esprit que ces répartitions sont les résultats d'études qui ont eu lieu entre 2005 et 2014 [6], puis ajustés en fonction du décompte des populations par pays. Il en résulte que les totaux de population sont fiables, mais la répartition de ces personnes à l'intérieur des territoires peut être périmée.

Ces données ont été créées à partir de multiples mesures lorsque le décompte de la population pour une sous-région n'a pas été considéré comme fiable. Les images satellites font d'ailleurs partie de ces mesures supplémentaires, ce qui pourrait avoir comme effet d'améliorer la capacité de notre réseau de neurones à apprendre de ces données (corrélation forte).

Ces cartes sont disponibles pour les années 2000, 2005, 2010, 2015 et 2020.

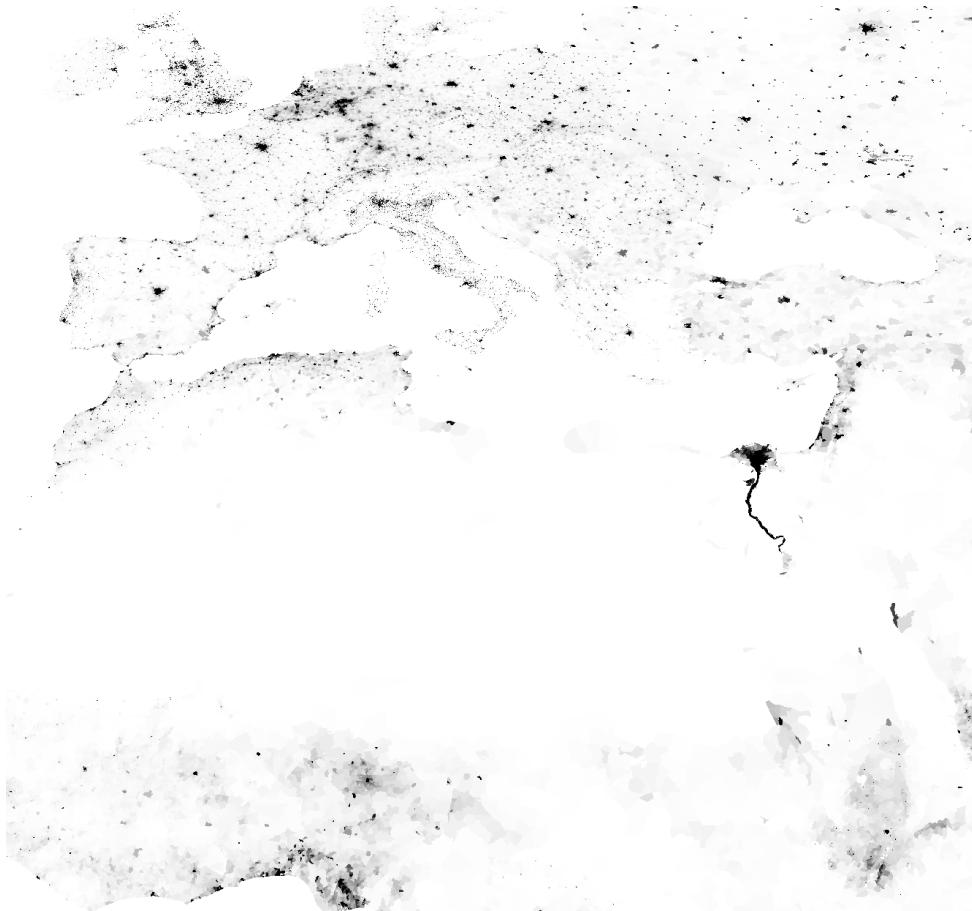


FIGURE 5.5 – Extrait de la grille de population [6] rendu avec QGIS. Le blanc indique une absence d'habitants, le noir indique au moins 1000 habitants par kilomètre carré.

5.1.3 Statistiques nationales

Les grilles de données globales sont difficiles à créer, il n'en existe donc pas de tous les types et toutes les dates. Afin de contourner ce problème, il y a des outils pour combiner des grilles (ici : l'image satellite) avec des données vectorielles telles que les frontières des pays. L'outil utilisé par le notebook `country_stats` est la librairie Python `rasterstats`.

Le jeu de données "Admin 0 - Countries" de Natural Earth [11] contient les frontières des pays actuelles, ainsi que des métadonnées sur chacun de ces pays (population estimée, indice de développement, différentes appellations et abréviations, etc.). En cas de conflits ou ambiguïtés politiques sur les frontières, c'est le pays qui contrôle le terrains qui est marqué comme souverain. Voici à quoi ressemblent les métadonnées pour la Tunisie :

('scalerank', 0), ('featurecla', 'Admin-0 country'), ('LABELRANK', 3.0), ('SOVEREIGNT', 'Tunisia'), ('SOV_A3', 'TUN'), ('ADM0_DIF', 0.0), ('LEVEL', 2.0), ('TYPE', 'Sovereign country'), ('ADMIN', 'Tunisia'), ('ADM0_A3', 'TUN'), ('GEOU_DIF', 0.0), ('GEOUNIT', 'Tunisia'), ('GU_A3', 'TUN'), ('SU_DIF', 0.0), ('SUBUNIT', 'Tunisia'), ('SU_A3', 'TUN'), ('BRK_DIFF', 0.0), ('NAME', 'Tunisia'), ('NAME_LONG', 'Tunisia'), ('BRK_A3', 'TUN'), ('BRK_NAME', 'Tunisia'), ('BRK_GROUP', None), ('ABBREV', 'Tun.'), ('POSTAL', 'TN'), ('FORMAL_EN', 'Republic of Tunisia'), ('FORMAL_FR', None), ('NAME_CIAWF', 'Tunisia'), ('NOTE_ADMIN0', None), ('NOTE_BRK', None), ('NAME_SORT', 'Tunisia'), ('NAME_ALT', None), ('MAPCOLOR7', 4.0), ('MAPCOLOR8', 3.0), ('MAPCOLOR9', 3.0), ('MAPCOLOR13', 2.0), ('POP_EST', 11403800.0), ('POP_RANK', 14.0), ('GDP_MD_EST', 130800.0), ('POP_YEAR', 2017.0), ('LASTCENSUS', 2004.0), ('GDP_YEAR', 2016.0), ('ECONOMY', '6. Developing region'), ('INCOME_GRP', '3. Upper middle income'), ('WIKIPEDIA', -99.0), ('FIPS_10_', 'TS'), ('ISO_A2', 'TN'), ('ISO_A3', 'TUN'), ('ISO_A3_EH', 'TUN'), ('ISO_N3', '788'), ('UN_A3', '788'), ('WB_A2', 'TN'), ('WB_A3', 'TUN'), ('WOE_ID', 23424967.0), ('WOE_ID_EH', 23424967.0), ('WOE_NOTE', 'Exact WOE match as country'), ('ADM0_A3_IS', 'TUN'), ('ADM0_A3_US', 'TUN'), ('ADM0_A3_UN', -99.0), ('ADM0_A3_WB', -99.0), ('CONTINENT', 'Africa'), ('REGION_UN', 'Africa'), ('SUBREGION', 'Northern Africa'), ('REGION_WB', 'Middle East & North Africa'), ('NAME_LEN', 7.0), ('LONG_LEN', 7.0), ('ABBREV_LEN', 4.0), ('TINY', -99.0), ('HOMEPART', 1.0), ('MIN_ZOOM', 0.0), ('MIN_LABEL', 3.0), ('MAX_LABEL', 8.0)

Le notebook `country_stats` superpose ces données vectorielles à l'image satellite pour ajouter à chaque pays sa luminosité moyenne ainsi que l'écart-type de luminosité par pixel. Puis ces données sont enregistrées dans le fichier `stats.pickle` pour être utilisées plus tard.

Grâce à cette information sur la luminosité perçue par pays, on peut faire un parallèle avec une grande variété de données nationales, telles que le produit intérieur brut (GDP), la consommation en énergie, le niveau de développement, l'indice économique, les émissions de CO₂, etc.

La source utilisée pour les données de population par pays et par année vient du site du Département des Affaires Économiques et Sociales des Nations Unies [13], plus précisément de la feuille Excel "Total Population - Both Sexes" de 2017.

La source de données pour le produit intérieur brut en USD par année est fournie par The World Bank [12] et téléchargée depuis http://data.un.org/Data.aspx?q=gdp&d=WDI&f=Indicator_Code%3aNY.GDP.MKTP.CD (visitée le 07.06.2018, dernière mise à jour le 12.10.2016).

Les données sur la consommation en électricité en millions de kWh par pays et par année sont tirées de <http://data.un.org/Data.aspx?d=EDATA&f=cmID%3AEL> (données issues de la Division Statistique des Nations Unies, mises à jour en janvier 2018). Le filtre "Electricity - Final energy consumption" a été utilisé pour ne garder que la consommation totale, sans les nombreux détails proposés par la base de données sur l'usage de l'électricité.

5.2 Recherche de corrélation

5.2.1 Statistiques nationales

La superposition de deux grilles (population et image satellite) de dimensions différentes n'est pas triviale, c'est pourquoi les premières données à avoir été mises en parallèle dans ce travail sont la population par pays [13], le produit intérieur brut [12], et la luminosité totale du pays

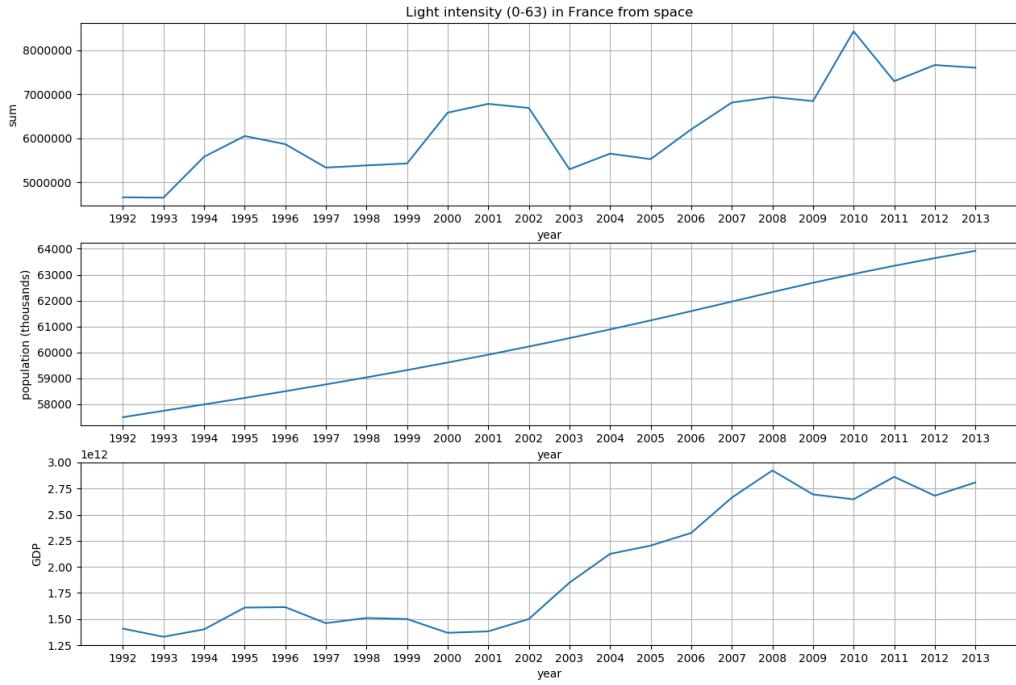


FIGURE 5.6 – Quantité de lumière perçue depuis l'espace, population et PIB de la France entre 1992 et 2013.

perçue depuis l'espace (extraite par script Python (notebook `country_stats`) à partir des vecteurs de frontières de Natural Earth [11] et des images de la NOAA [8]).

Pour chaque année disponible de 1992 à 2013, il est donc possible de dessiner pour la majorité des pays les histogrammes de luminosité, population et PIB (exemples en figures 5.6, 5.7, 5.8 et 5.9). Cette représentation n'est pas idéale, car l'échelle n'est pas constante entre les pays. Ce qui apparaît comme une grande variation de luminosité peut ne pas en être du tout. Certains pays ne portent pas exactement le même nom dans tous les jeux de données, ils sont donc mis de côté dans les résultats.

La figure 5.6 est typique des résultats obtenus. Elle laisse présager une forte corrélation entre les trois données observées. Cependant, la figure 5.7 soulève déjà des doutes : la population croît de façon linéaire, mais le PIB et la lumière perçue augmentent de plus en plus vite. Les graphes du Japon (figure 5.8) semblent indiquer que la luminosité suit le PIB également lorsque ce dernier diminue. Enfin, le cas de l'Arménie (figure 5.9) nous pousse à croire que le PIB a une corrélation plus forte que la population avec la luminosité.

Jusqu'ici, rien n'est confirmé. Nous ne possédons que d'observations par pays sur des échelles variables, ce qui représente un niveau de détail très bas. Sans parler du fait que 320 pays sur 21 années ne sont pas une quantité de données acceptable pour entraîner un réseau de neurones (6720 observations).

Il est également possible de générer des nuages de points comparant la lumière émise et, au choix, l'énergie consommée, la population du pays ou le PIB, et de colorer ces points par indice de développement économique (tiré des données vectorielles Natural Earth [11]). Tous ces graphes sont générés grâce au notebook `country_stats`.

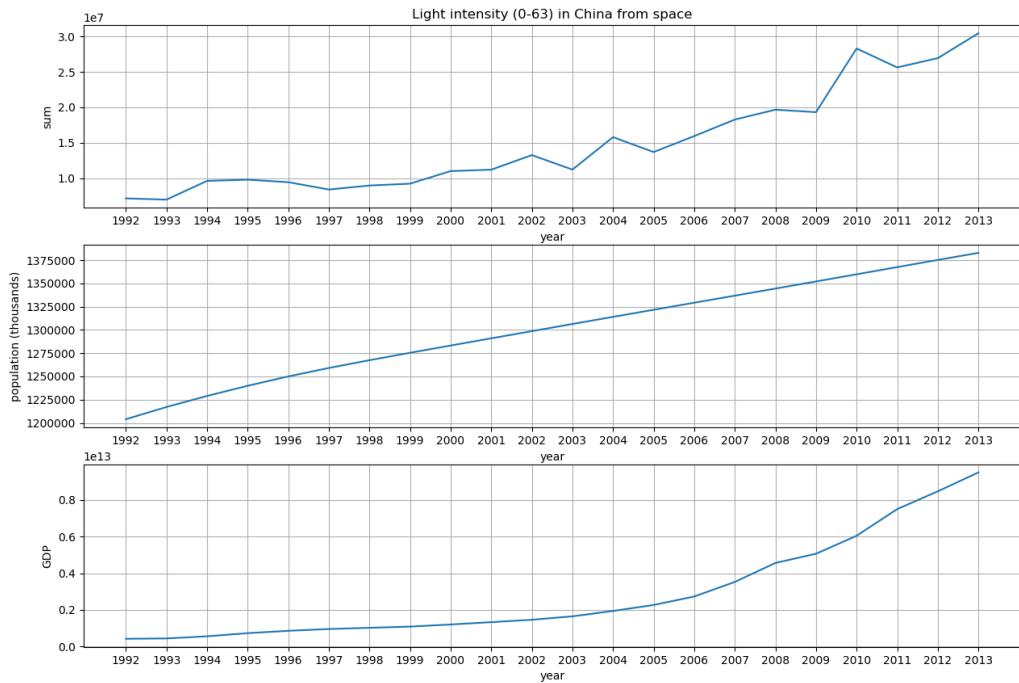


FIGURE 5.7 – Quantité de lumière perçue depuis l'espace, population et PIB de la Chine entre 1992 et 2013.

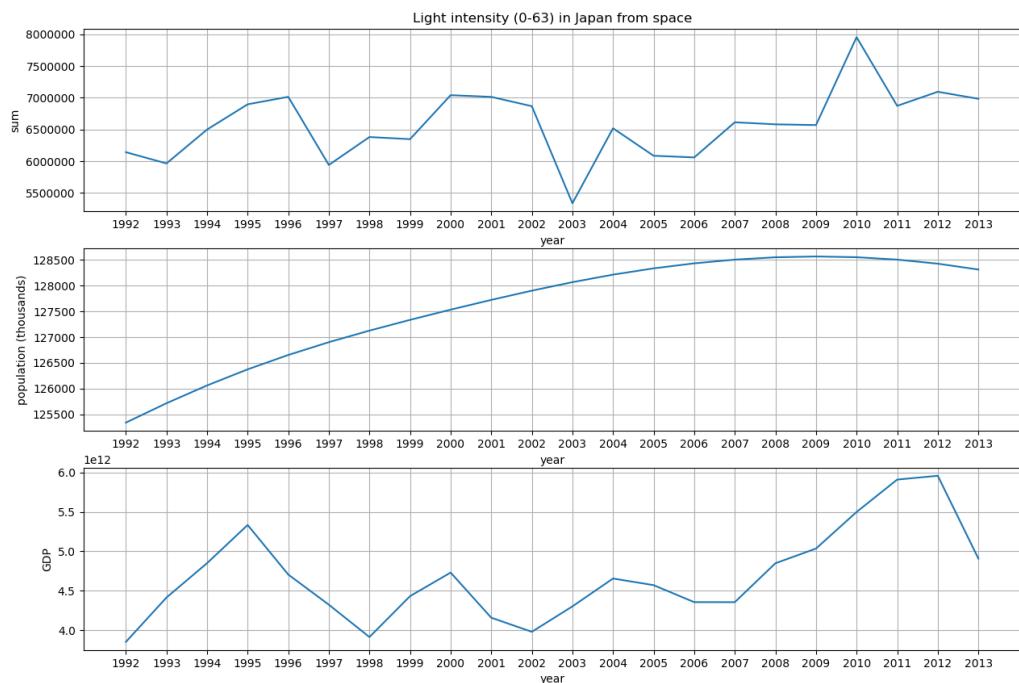


FIGURE 5.8 – Quantité de lumière perçue depuis l'espace, population et PIB du Japon entre 1992 et 2013.

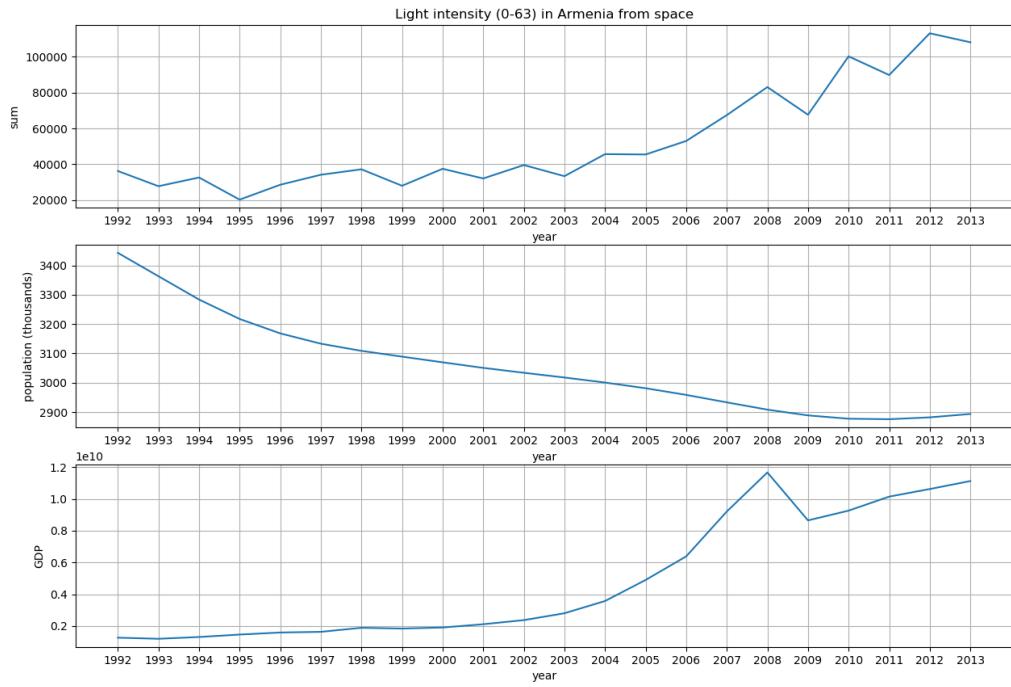


FIGURE 5.9 – Quantité de lumière perçue depuis l'espace, population et PIB de l'Arménie entre 1992 et 2013.

Le premier de ces nuages de points compare la luminosité et la population de 2013 pour chaque pays (figure 5.10). On observe immédiatement qu'il existe une corrélation. Son coefficient de Pearson est 0.56. Il est très important de constater que cette corrélation n'apparaît que lorsque l'échelle est logarithmique, il faudra donc retenir que cette relation n'est pas directement linéaire. On ne peut pas encore déterminer s'il y a une relation directe, ou si ces deux variables sont simplement corrélées à la taille du territoire. En effet, on ne fait que sommer la population et la quantité de lumière émise, on ne calcule pas de moyennes par pays. Sans surprise, on voit également que les pays ayant un index économique élevé émettent plus de lumière que d'autres pays à population équivalente.

Le deuxième graphe généré est beaucoup plus intéressant, car il compare le produit intérieur brut de chaque pays avec leur émission de lumière (figure 5.11). On observe une forte corrélation, dont le coefficient de Pearson est 0.819. Naturellement, l'indice de développement économique est très corrélé avec le PIB (GDP en Anglais).

Une autre comparaison est possible avec la consommation en électricité, illustrée en figure 5.12. La forte corrélation (coefficient de 0.81) entre émission de lumière et consommation d'énergie n'est pas une surprise, mais on remarque un phénomène intéressant avec l'index économique des pays. Les pays en voie de développement ont tendance à émettre beaucoup plus de lumière pour une consommation d'électricité équivalente aux pays développés. On peut spéculer sur les causes d'une telle différence (data centers, chauffages, etc.) mais ce n'est pas l'objet de ce travail.

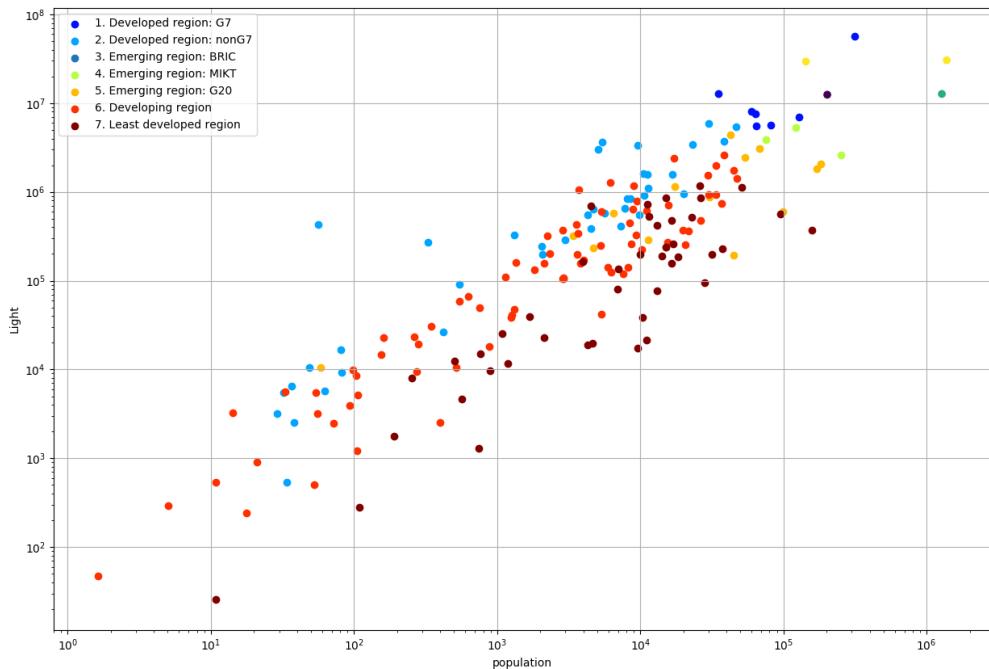


FIGURE 5.10 – Pays placés par population et luminosité totale émise sur une échelle logarithmique en 2013, colorés par indice économique.

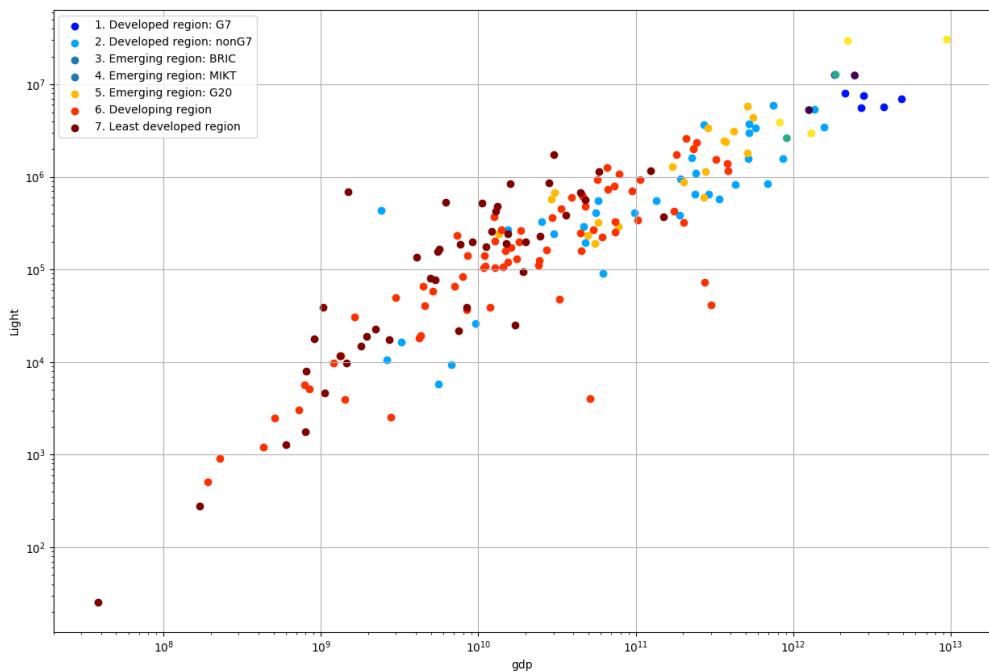


FIGURE 5.11 – Pays placés par PIB en USD et luminosité totale émise sur une échelle logarithmique en 2013, colorés par indice économique.

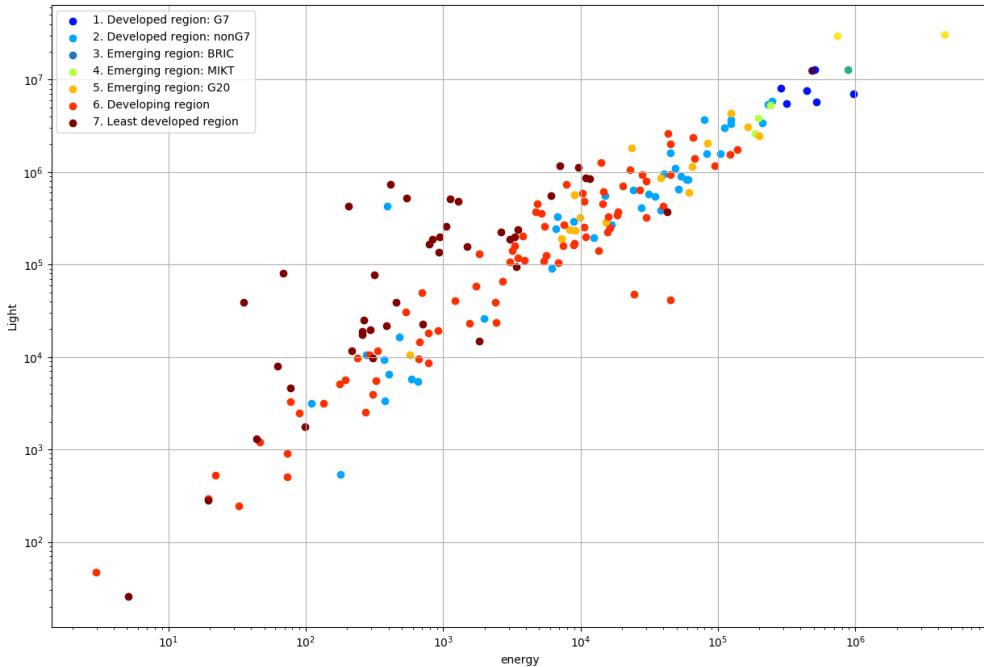


FIGURE 5.12 – Pays placés par consommation en électricité en millions de kWh et luminosité totale émise sur une échelle logarithmique en 2013, colorés par indice économique.

5.2.2 Grille de population

Après avoir tenté d'écrire un script Python pour superposer deux grilles de dimensions différentes, il s'est avéré que l'application QGIS est capable d'effectuer cette opération. Il suffit de :

- Ouvrir les grilles avec QGIS : population et image satellite. Elles doivent apparaître dans la liste des couches du projet.
- Dans la barre d'outils, choisir "Raster" → Divers → Fusionner.
- Cocher l'option "Placer chaque fichier en entrée dans une bande séparée." puis lancer la fusion "Run in Background".

La grille résultante contient donc environ 800 millions de pixels (possédant chacun une valeur de population et de luminosité) qui sont potentiellement autant de données d'entraînement pour chaque année disponible (2000, 2005, 2010, qui correspondent aux dates des grilles de population et de images DMSP-OLS [8]). Après avoir ajusté ses propriétés d'affichage et inversé les couleurs, on obtient la figure 5.13. Le bleu ciel correspond à la lumière visible depuis l'espace, le rose la population (sur une échelle de 1 à 1000 habitants par km²). Le bleu foncé correspond au chevauchement des deux couleurs. On peut déjà observer que, si la luminosité ne suit pas l'évolution de la population dans le temps à l'échelle d'un pays, elle est tout de même concentrée géographiquement sur les points les plus peuplés.

La superposition des grilles de population (1 pixel par km) avec une image satellite issue de VIIRS [7] dont la résolution est 4 fois plus élevée pose problème. Les outils de traitement de cartes ont tendance à augmenter la résolution de la grille de population en utilisant la méthode du



FIGURE 5.13 – Superposition de l'image satellite nocturne (bleu ciel) et de la grille de population (rose).

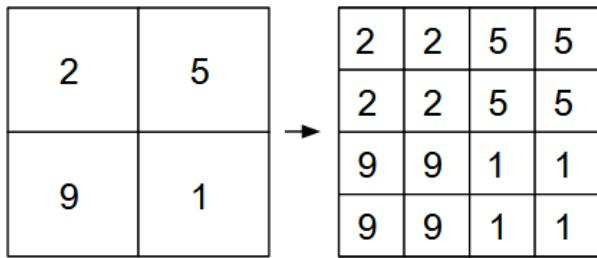
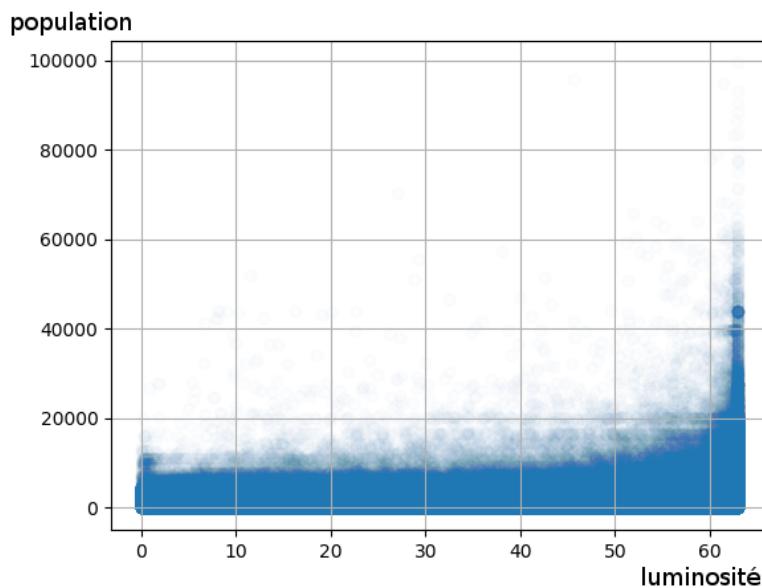


FIGURE 5.14 – Effet de l'augmentation de la résolution de la grille de population lors de la fusion

FIGURE 5.15 – Valeur de luminosité et de population mondiale pour chaque km² (année 2000)

plus proche voisin. Chaque pixel donne donc correctement la densité de population par kilomètre carré, mais la population totale est multipliée par 4. La figure 5.14 illustre ce problème. Pour corriger cela, il est nécessaire d'ajuster la population par pixel en la divisant par 4. C'est le rôle du script `gleam/rescale_pop.py` (attention à ajuster les valeurs des constantes avant de l'exécuter).

Si l'on pose chaque pixel sur un nuage de points dont les axes sont la luminosité et la population, on obtient la figure 5.15. La figure 5.16 représente uniquement le Brésil en 2015. On constate qu'il sera très difficile de deviner la valeur de population à partir de l'intensité lumineuse d'un seul pixel. Cela nous oriente donc plutôt dans la direction d'un réseau de neurones à convolutions, qui considère les pixels par groupes. Le coefficient de corrélation entre luminosité et population au Brésil est 0.72625. On considérera donc qu'une prédiction est plutôt bonne si son coefficient de corrélation atteint cette valeur lorsqu'elle est comparée à la grille de population.

Enfin, les données vectorielles de Natural Earth [11] peuvent nous permettre d'isoler précisément un pays du reste de la grille :

- Ouvrir la grille et le fichier vectoriel des pays avec QGIS dans le même projet.
- Cliquer sur la couche vectorielle, puis sélectionner sur la carte le pays à isoler (avec l'outil de sélection d'entités).
- Dans la barre d'outils, sous "Raster", choisir "Extraction" puis "Découper un raster selon une couche masque".

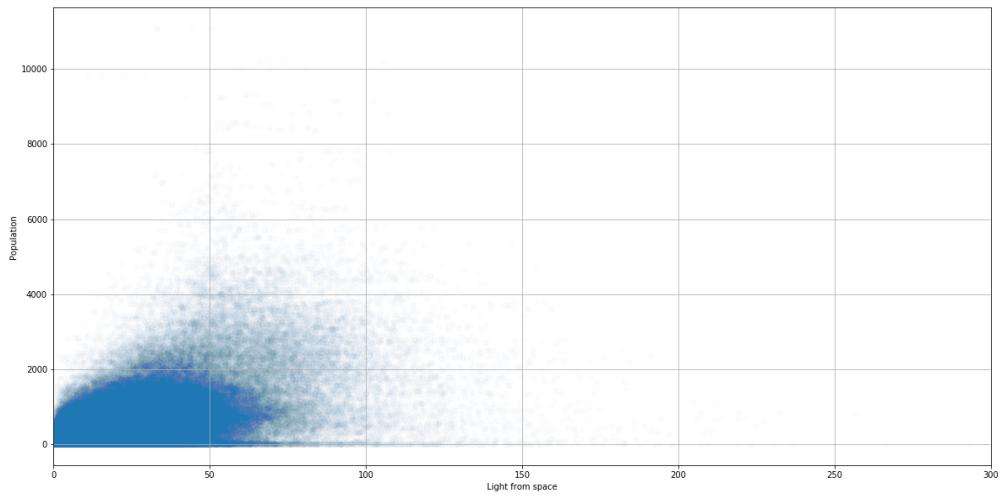


FIGURE 5.16 – Valeur de luminosité et de population pour chaque 0.25km^2 du Brésil (année 2015)

- Définir la grille comme couche source, et la couche vectorielle comme masque.
- Cocher "Entité(s) sélectionnée(s) uniquement".
- Lancer l'extraction "Run in Background".

6 Modèle

6.1 Environnement de développement

Python est un langage largement utilisé pour manipuler des données et faire de l'apprentissage automatique. Il dispose de librairies optimisées précisément pour cette tâche, dont Keras.

Keras est une librairie Python open source permettant le prototypage rapide de réseaux de neurones et peut fonctionner au-dessus de TensorFlow (open source et développé par Google), CNTK ou Theano. Une fonctionnalité très attractive de TensorFlow est l'exploitation automatique du processeur graphique s'il est disponible. En effet, ces processeurs sont très performants pour l'apprentissage automatique, car c'est un travail hautement parallélisable.

Afin d'activer l'utilisation d'un GPU Nvidia par Tensorflow sur Windows 10, il faut installer la librairie tensorflow-gpu, le CUDA Toolkit de Nvidia en version 9.0 et le SDK cuDNN (*CUDA Deep Neural Network library*) de Nvidia en version 7.0 (instructions détaillées sur <https://www.tensorflow.org/install/>).

6.2 Réseau de neurones

Pour commencer à développer le réseau de neurones, on commencera par tenter de prédire la population d'une petite région à partir de l'image satellite de l'année 2005, en entraînant la machine sur celle de 2000 [8] et la grille de population correspondante [6].

Comme on a pu observer que la prédiction à partir d'un pixel isolé n'est pas réaliste, on aimerait prendre en compte les valeurs des pixels voisins. En effet, s'il y a beaucoup de pixels proches illuminés, il y a de meilleures chances pour que la région soit densément peuplée que si un seul pixel est illuminé. De plus, si le réseau de neurones peut reconnaître les formes, il sera en mesure de différencier le centre et la périphérie d'une ville, ainsi que des routes ou villages isolés.

Le système qui répond à ces exigences est le réseau de neurones à convolutions. Cependant, alors que l'usage habituel d'un tel réseau sert à la classification de données, on a besoin ici d'obtenir un nombre réel. Il s'agit d'adapter le système pour faire de la régression, ce que l'on fait lorsqu'on compte le nombre de voitures sur un parking par exemple. Les excellents articles *Regression Tutorial with the Keras Deep Learning Library in Python* [5], *Crash Course in Convolutional Neural Networks for Machine Learning* [1], *Handwritten Digit Recognition using Convolutional Neural Networks in Python with Keras* [3], *Object Recognition with Convolutional Neural Networks in the Keras Deep Learning Library* [4] et *Evaluate the Performance Of Deep Learning Models in Keras* [2] de Jason Brownlee ont été d'une grande aide pour la compréhension des bonnes pratiques et l'utilisation de Keras.

Avant de commencer l'entraînement, on découpe l'image satellite nocturne en tuiles de 32 sur 32 pixels (1024 km^2). Chacune de ces tuiles sera considérée comme une observation à donner en entrée du modèle. La sortie sera une valeur réelle correspondant au nombre d'habitants dans la zone donnée en entrée.

6.3 Première topologie

La topologie du réseau (figure 6.1) s'inspire grossièrement du travail effectué par l'Arnhold Institute for Global Health [9], qui consiste à estimer les populations de petites régions à partir d'images satellite de jour *et* de nuit. La première version de notre réseau de neurones en utilise une variante très allégée.

Les couches sont définies comme suit :

- Convolution de 32 filtres, chaque kernel fait 3x3 pixels, fonction d'activation ReLU,
- Max pooling pour diviser par 2 la largeur et la hauteur de la tuile,
- Convolution identique à la première couche,
- Flatten,
- Dropout de 20% pour réduire les chances d'overfitting,
- Couche dense de 32 neurones,
- Couche dense d'un seul neurone, qui correspond à la sortie du modèle.

L'optimiseur utilisé est Adam. C'est généralement un bon choix par défaut grâce à sa performance en terme de vitesse de calcul, et sa capacité à répondre aux besoins d'un grand nombre de problèmes. Le *learning rate* choisi (après quelques essais) est 0.001. Enfin, la fonction objectif à optimiser est la moyenne des erreurs au carré.

6.4 Premiers résultats

Les résultats suivants ont été obtenus en entraînant le modèle sur une région couvrant l'Europe de l'Ouest et l'Afrique du Nord en l'an 2000. La phase de test est effectuée sur l'Amérique du Nord et une partie de l'Amérique du Sud en l'an 2005.

Les figures 6.2 et 6.3 présentent respectivement l'évolution des moyennes des erreurs absolues et des erreurs au carré pendant l'entraînement du modèle sur 100 itérations.

Les résultats du test sur l'Amérique en 2005 sont plutôt médiocres. La fonction objectif (moyenne des erreurs au carré) vaut 5381938110.04, et la moyenne des erreurs absolues vaut 26599.64. Ce qui veut dire que sur une région de 1024 km^2 , la prédiction du modèle se trompe en moyenne de 26599.64 habitants. C'est significativement plus mauvais que les résultats obtenus sur le jeu d'entraînement. Cela signifie que le modèle apprend, car l'erreur diminue lors de l'entraînement, mais il n'apprend rien de généralisable. C'est ce que l'on appelle le surapprentissage. De plus, une erreur moyenne de 26599.64 par tuiles représente une erreur moyenne d'environ 26 habitants par kilomètre carré. À titre de comparaison, la densité de population du monde est d'environ $7000000000 / 500000000 = 14$ habitants par kilomètre carré, océans compris.

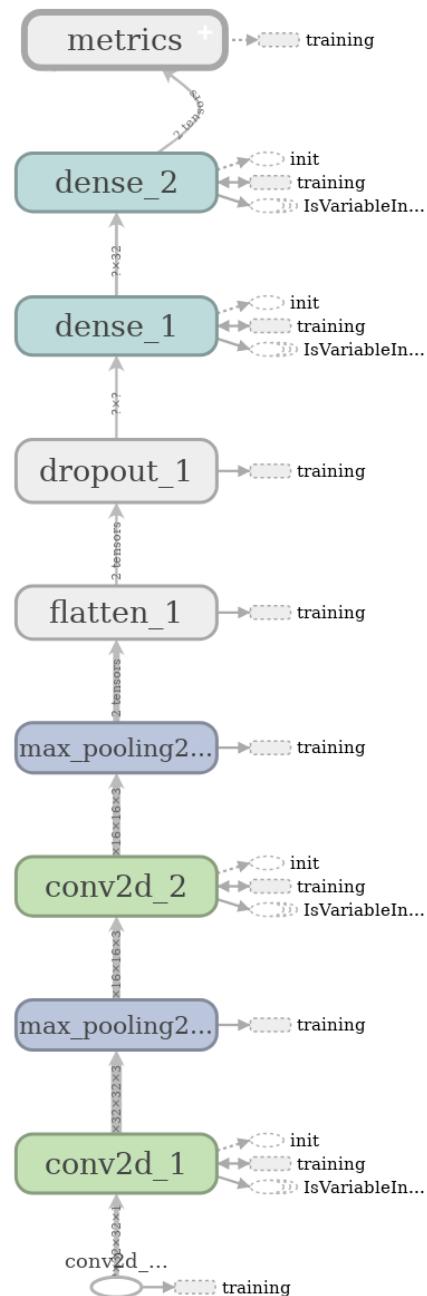


FIGURE 6.1 – Topologie de la toute première version du réseau de neurones.

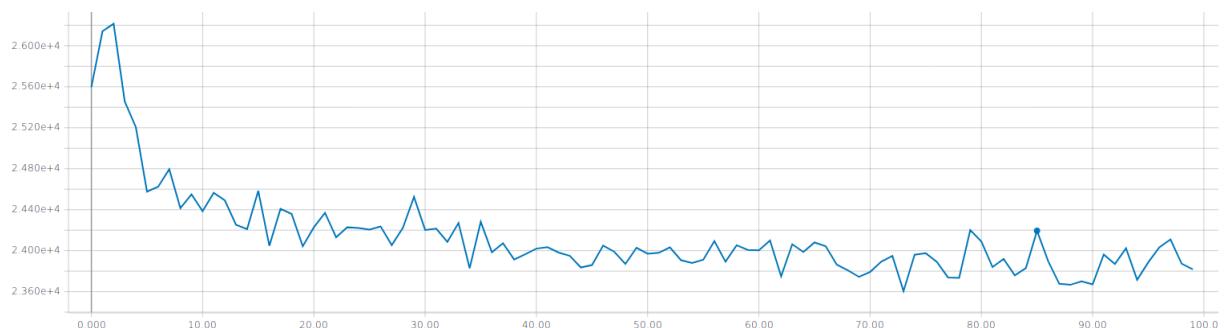


FIGURE 6.2 – Moyennes des erreurs absolues sur 100 itérations.

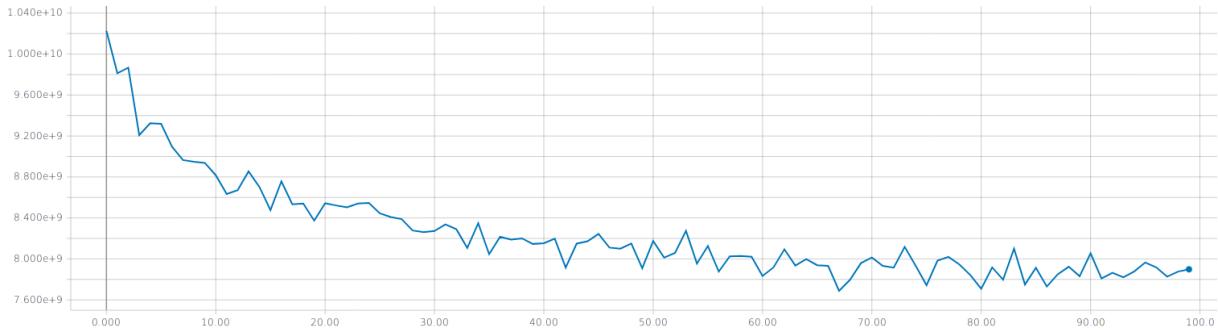


FIGURE 6.3 – Moyennes des erreurs au carré (fonction objectif) sur 100 itérations.

6.5 Améliorations

6.5.1 Prétraitement

La plus importante amélioration en performance que l'on peut obtenir sur un réseau de neurones vient d'un prétraitement adapté. La façon dont les informations sont présentées au modèle influence grandement sa capacité à apprendre. Cet aspect n'a pas été travaillé pour le premier modèle. Voici donc les améliorations apportées.

Premièrement, la Terre est recouverte à plus de 70% d'eau. Cette surface n'est pas occupée par des habitations. Le jeu de données comporte donc une grande majorité d'observations superflues qui peuvent interférer avec l'apprentissage de caractéristiques plus importantes. Le processus de création des observations de 32 sur 32 pixels doit donc exclure les éléments vides. On filtre donc les tuiles sans population de l'entraînement et la validation.

De plus, sans même parler de l'influence de l'économie d'un pays sur ses émissions de lumière, la qualité des données de population est très inégale entre les pays. La figure 6.4 en est un exemple. Sur la grande majorité des pays du monde, et même de l'Europe, la grille de population ne propose qu'une seule estimation de densité pour de grandes régions. Cela devient particulièrement problématique lorsqu'une observation contient une ville isolée émettant de la lumière, mais qu'elle se trouve au milieu d'une grande région dont la densité de population est basse. Ces cas sont nombreux et réduisent la corrélation entre population et luminosité que le modèle doit exploiter pour apprendre. Afin de gagner en performance, on peut découper la carte en isolant un pays dont les données de population sont de bonne qualité. C'est faisable avec QGIS par exemple, et un fichier vectoriel pour les frontières tel que celui de Natural Earth [11]. Tous les tests dans le but d'améliorer la topologie du réseau de neurones seront faits sur le Brésil uniquement (bonne qualité, grande surface).

Cependant, restreindre l'entraînement à un pays réduit la taille du jeu de données et augmente le risque de surapprentissage. Pour augmenter la quantité d'observations, chaque tuile de 32 sur 32 pixels est décalée de seulement 8 pixels au lieu de 32. Ainsi, aucune observation n'est identique, mais elles se superposent partiellement. Il est donc judicieux de ne pas utiliser cette méthode lors de la validation croisée.

Enfin, réduire la taille des tuiles a donné de moins bons résultats lors de la validation, leur taille restera donc de 32 sur 32 pixels. L'image satellite et la grille de populations utilisée [6] seront celles de 2015 afin d'exploiter les images de plus haute résolution (500m par pixel) de la NOAA [7].

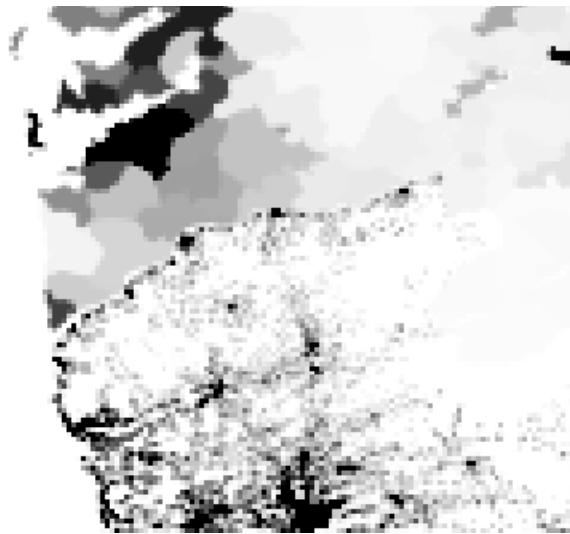


FIGURE 6.4 – Grille de population pour la frontière nord entre l'Espagne et le Portugal illustrant les différences de qualité des données de population d'un pays à l'autre.

6.5.2 Topologie et paramètres

Plus de 50 modèles ont été générés dans le but de tester différents ajustements sur le réseau de neurones et ses paramètres. Ces tests ont été effectués en utilisant la validation croisée 4-fold sur au moins 150 itérations. Le jeu de données d'entraînement est le Brésil en 2015, les tuiles ne se superposent pas (décalage de 32 pixels entre chaque observation de 32 sur 32). Voici un résumé non exhaustif de ces tests.

- Ajustements de la taille des filtres des convolutions,
- Ajouts/suppressions de couches de convolution et de max pooling,
- Ajouts/suppressions de couches denses et de dropout,
- Ajustements de la force du max pooling,
- Ajustements du learning rate,
- Implémentation du learning rate dégressif,
- Pondération des observations selon la population présente (non concluant),
- Ajustements du nombre de neurones dans la couche dense,
- Ajustements des valeurs de dropout,
- Passage du max pooling à l'average pooling, car lorsque l'on réduit la résolution des grilles on veut sommer les valeurs plutôt que de garder la valeur maximale des pixels. L'average pooling remplit ce rôle, car la moyenne des pixels est la somme divisée par une constante dans le cas du pooling.

La section 6.6 détaille la topologie et les paramètres retenus après ces tests.

6.5.3 Post-traitement

Les prédictions ne peuvent être faites que sur des images satellites de même échelle (en mètres par pixel) que celle qui a servi à entraîner le réseau. Le résultat de la prédiction est une estimation de la population présente sur l'observation en entrée du modèle, ce qui réduit fortement la résolution de l'image en sortie. Cependant, grâce à l'article *Equitable development through deep learning* [9] et l'exploration préalable des données, on sait que la répartition de

population est en relation exponentielle avec la répartition de la luminosité.

Cette constatation nous permet de répartir approximativement la population prédictive par le modèle pour recréer une tuile de dimension identique à l'observation en entrée du modèle, c'est-à-dire 1024 pixels. On peut valider cette répartition en calculant le coefficient de corrélation entre la prédiction et la donnée d'entraînement. Avec le modèle entraîné sur le Brésil pour 2015, en comparant la grille de population à la prédiction de la population du Brésil en 2015, on obtient les coefficients de corrélation :

- 0.60771883 avec une distribution pondérée linéairement à la luminosité des pixels de la tuile,
- 0.69339848 avec une distribution exponentielle,
- 0.63450044 avec une distribution logarithmique,
- 0.65867300 avec une distribution quadratique.

Ces résultats s'alignent avec l'hypothèse que la distribution exponentielle est la plus adaptée.

6.6 Réseau définitif

TABLE 6.1 – Topologie du réseau en version définitive

Couche (nb de neurones)	Dimensions de la sortie	Nb de paramètres
Convolution (64)	$32 \times 32 \times 64$	640
Average pooling	$16 \times 16 \times 64$	0
Convolution (128)	$16 \times 16 \times 128$	73856
Average pooling	$8 \times 8 \times 128$	0
Convolution (256)	$8 \times 8 \times 256$	295168
Average pooling	$4 \times 4 \times 256$	0
Flatten	4096	0
Dropout 50%	4096	0
Dense (128)	128	524416
Dense (1)	1	129
Total		894209

La topologie du réseau retenu est détaillée sur la table 6.1. Les paramètres supplémentaires sont les suivants :

- La taille des filtres est 3 sur 3 pixels pour toutes les couches de convolution.
- Chaque average pooling divise la largeur et la hauteur de l'image en entrée de la couche par 2.
- Le dropout abandonne 50% des neurones lors des étapes d'entraînement pour éviter le surapprentissage.
- L'ajustement des paramètres (rétropropagation) utilise l'algorithme Adam, et ne s'effectue que toutes les 1024 mesures. Cela permet de paralléliser les calculs efficacement pour un processeur graphique et réduit le nombre de rétropropagations à effectuer.
- La fonction objectif est l'erreur quadratique moyenne. Cela permet de réduire la variance des erreurs ainsi que l'erreur absolue.
- Le learning rate commence à 0.02. Il est divisé par 2 lorsqu'il n'y a pas eu d'amélioration de la fonction objectif dans les 20 dernières itérations. Il ne descend pas plus bas que 0.0000001.
- Si la fonction objectif ne s'améliore pas pendant 200 itérations, l'entraînement s'arrête.

- Le nombre maximal d'itérations est fixé à une valeur très élevée de sorte à ne pas interrompre l'entraînement tant que la fonction objectif s'améliore.
- L'ordre de traitement des observations est randomisé à chaque itération pour ne pas grouper les observations par région.

6.7 Résultats

6.7.1 Validation croisée

Le résultat de la validation croisée 4-fold lors de l'entraînement sur le Brésil en 2015 est présenté dans la table 6.2. Le prétraitement utilisé pour la validation croisée ne génère pas d'observations qui se superposent (chaque tuile est décalée de 32 pixels de la précédente).

TABLE 6.2 – Résultat de la validation croisée lors de l'entraînement sur le Brésil

Mesure	Valeur moyenne	Écart-type
Erreur quadratique moyenne	234181700.80	99980235.55
Erreur absolue moyenne	2021.24	194.84
Erreur absolue moyenne par km ²	7.9	0.76
Somme des erreurs absolues	19786348.86	1906802.05

La validation étant 4-fold, elle consiste en 4 entraînements indépendants dont les données de validation sont toujours différentes. L'écart-type indiqué dans les résultats est l'écart-type entre les 4 mesures obtenues.

6.7.2 Évaluation des prédictions

Afin d'évaluer la qualité des prédictions, on compare chaque prédiction avec la grille de population [6] utilisée lors de l'entraînement, pixel par pixel, grâce au notebook `rastercomparator`. Cette comparaison ignore les pixels qui ont une valeur égale à 0. Cela permet d'éliminer les océans et les territoires hors frontières du calcul du coefficient de corrélation et de l'erreur moyenne. Sans cela, l'erreur serait anormalement basse et non représentative.

Les pays choisis pour les mesures sont des pays pour lesquels la grille de population [6] est la plus précise : Brésil, USA, Portugal, Italie, Iles de Bretagne. Les modèles testés sont ceux qui ont été entraînés sur le Brésil ou les USA en 2015 à cause de leur grande superficie. On tente de mettre en relation des pays dont le développement économique est vaguement proche. Les résultats sont présentés dans les tables 6.3 à 6.8.

Afin de donner des pistes pour comprendre la source des erreurs, on peut soustraire les données prédites à la grille de population. Un extrait du résultat est présenté en figure 6.5.

6.8 Discussion

La validation croisée nous donne une idée de la performance du modèle sans l'algorithme de prédiction qui reconstruit la grille complète. Le résultat est donc un seul nombre pour la population sur une tuile de 32 sur 32 pixels, ce qui représente dans ce cas une zone de 256 km².

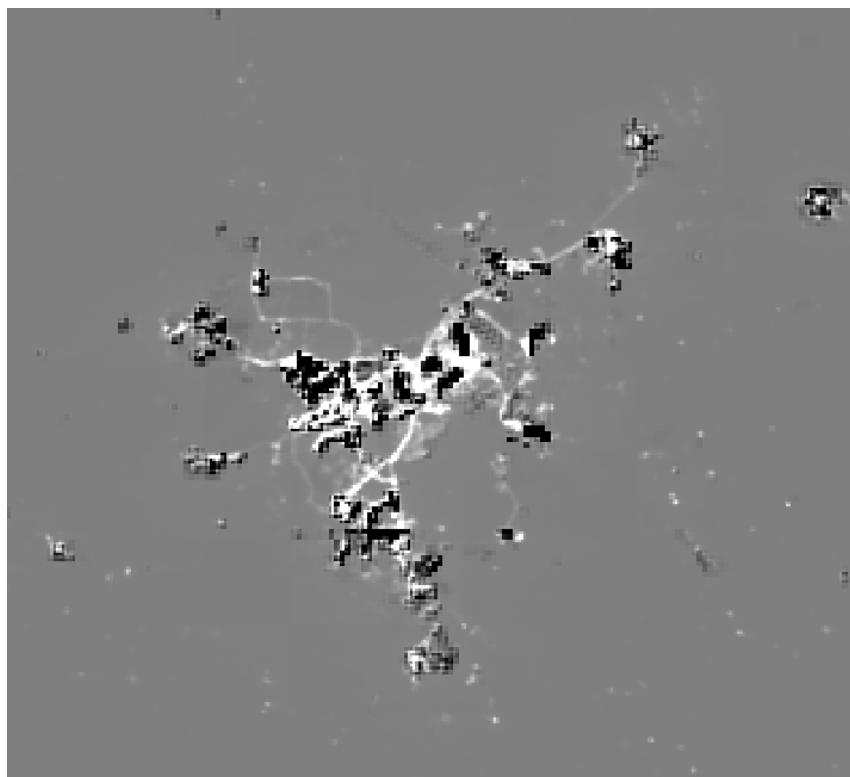


FIGURE 6.5 – Résultat de la soustraction de la grille de population SEDAC [6] par la prédiction du modèle entraîné sur le Brésil en 2015. La ville représentée est Brasilia. Le blanc signifie une surestimation de la population de 500, le noir signifie une sous-estimation de la population de 500 par pixel (0.25 km^2).

TABLE 6.3 – Modèle entraîné sur le Brésil en 2015, prédiction du Brésil en 2015

Mesure	Valeur
Population totale réelle	204777020
Population totale prédite	203754690
Augmentation du total	-1022336
Augmentation du total (%)	-0.49924%
Erreur absolue moyenne par km^2	16.80172
Corrélation de Pearson	0.69289

TABLE 6.4 – Modèle entraîné sur le Brésil en 2015, prédiction du Portugal en 2015

Mesure	Valeur
Population totale réelle	10074708
Population totale prédite	25155508
Augmentation du total	15080800
Augmentation du total (%)	149.68969%
Erreur absolue moyenne par km^2	139.74304
Corrélation de Pearson	0.77799

TABLE 6.5 – Modèle entraîné sur les USA en 2015, prédiction des USA en 2015

Mesure	Valeur
Population totale réelle	319627900
Population totale prédictive	306326000
Augmentation du total	-13301888
Augmentation du total (%)	-4.16167%
Erreur absolue moyenne par km ²	26.88018
Corrélation de Pearson	0.66294

TABLE 6.6 – Modèle entraîné sur les USA en 2015, prédiction de l'Italie en 2015

Mesure	Valeur
Population totale réelle	58895960
Population totale prédictive	37188820
Augmentation du total	-21707140
Augmentation du total (%)	-36.85675%
Erreur absolue moyenne par km ²	117.458312
Corrélation de Pearson	0.63246

TABLE 6.7 – Modèle entraîné sur les USA en 2015, prédiction du Portugal en 2015

Mesure	Valeur
Population totale réelle	10074708
Population totale prédictive	8911347
Augmentation du total	-1163361
Augmentation du total (%)	-11.54734%
Erreur absolue moyenne par km ²	59.62044
Corrélation de Pearson	0.71329

TABLE 6.8 – Modèle entraîné sur les USA en 2015, prédiction des Iles de Bretagne en 2015

Mesure	Valeur
Population totale réelle	68446010
Population totale prédictive	35654788
Augmentation du total	-32791220
Augmentation du total (%)	-47.90815%
Erreur absolue moyenne par km ²	84.077484
Corrélation de Pearson	0.59551

Il est important de rappeler que, lors de la validation croisée, afin que toutes les observations ne se superposent pas, le prétraitement ne génère que le quart des tuiles qui sont à disposition lors de l'entraînement complet.

Selon The World Bank [12], en 2015 le Brésil possédait 24.642 habitants par km^2 . Notre modèle se trompe en moyenne de 7.9 habitants par km^2 . Bien que cette erreur ne soit pas négligeable, ce résultat paraît très satisfaisant si l'on considère le jeu de données réduit, et le fait que la grille servant à l'entraînement utilise une répartition de la population datant de 2005 à 2014 [6]. Nous n'avons malheureusement pas de moyens de savoir dans quelle proportion cette erreur est due au jeu de données vieillissant ou à la performance du réseau de neurones.

Passons aux cartes générées en pleine résolution après la prédiction par un modèle bien entraîné (mais sans validation croisée). Le résultat de ces mesures (tables 6.3 à 6.8) est fortement dépendant de la façon dont on répartit la population prédictive à l'intérieur des tuiles. Cela explique l'erreur par km^2 de la table 6.3 plus élevée que lors de la validation croisée.

Comparer la population totale prédictive avec la réalité peut indiquer si deux pays sont comparables, c'est-à-dire si un pays peut décemment servir de données d'entraînement pour une prédiction sur l'autre pays. Cependant, la mesure n'est pas bonne pour évaluer la qualité d'une prédiction. Le modèle cherche à estimer correctement la population d'une région de 256 km^2 indépendamment des autres observations. Il ne fait aucun rapprochement avec la population totale du pays, ce qui fait que de petites erreurs peuvent s'accumuler.

Le modèle entraîné sur le Brésil se trompe en moyenne de 34.19 habitants par pixel, soit 139.74 par km^2 , lorsqu'il est appliqué à l'image satellite du Portugal. C'est moins bon que le modèle entraîné sur les USA, qui se trompe de 59.62 par km^2 . La densité de population du Portugal était de 113.072 en 2015 selon The World Bank [12]. Le coefficient de corrélation plus élevé pour la prédiction du modèle entraîné sur le Brésil semble indiquer qu'il produit cependant moins de grosses erreurs. Cet exemple illustre l'importance de bien choisir les pays à comparer.

De manière générale, les coefficients de corrélation sont plutôt bons. Nous avons trouvé qu'il valait 0.72625 lors de la comparaison entre l'image satellite du Brésil et la grille de population de validation. Le fait que les coefficients trouvés tournent autour de 0.7 est donc très encourageant.

Un coefficient proche de 1 ou -1 indique que les valeurs prédictives sont proportionnelles aux valeurs de validation. Il est donc peu affecté par les différences d'intensité lumineuse par habitant entre le pays d'entraînement et de validation. C'est la raison pour laquelle le modèle entraîné sur le Brésil possède un si bon coefficient de corrélation avec la grille de validation lorsqu'il est appliqué au Portugal. Cette prédiction n'est pas entièrement mauvaise, mais elle nécessiterait d'être remise à l'échelle pour donner des valeurs exploitables. Si l'on divise la population prédictive par 2.5 pour adapter la somme des pixels prédictifs à la population du pays, le coefficient de corrélation reste identique, mais l'erreur absolue par km^2 descend à 58.7317. Un score rivalisant avec celui obtenu par le modèle entraîné sur les USA.

La figure 6.5 nous montre où sont les grosses erreurs. On aperçoit les routes en blanc. En effet, celles-ci émettent beaucoup de lumière sans confirmer pour autant la présence d'habitants au bord de ces routes, ce qui a pour conséquence une surestimation de la population à ces endroits.

L'autre anomalie qui saute aux yeux est le bruit observé en centre-ville. On trouve des pixels très noirs à côté d'autres très clairs. Ce n'est pas dû au hasard. Si l'on compare ces taches à la carte de Brasilia, on remarque que les quartiers où l'on surestime la population sont commerciaux ou riches, alors que ceux où l'on sous-estime la population sont pauvres et densément peuplés. On touche donc à la limite des informations que les images satellites nocturnes peuvent nous

offrir.

6.9 Prédictions non validées

Pour exploiter ce modèle, on aimeraient générer des cartes de population pour des pays qui n'en disposent pas. La Colombie est notamment intéressante, car l'apaisement du conflit contre les Forces armées révolutionnaires de Colombie pourrait provoquer un déplacement de la population hors des villes et vers la forêt Amazonienne. Bien que l'on n'en perçoive pas encore l'impact sur la luminosité émise par les habitations dans ces régions, il serait très intéressant d'observer des changements sur les années qui viennent. Notre modèle serait en mesure d'approximativement quantifier les migrations.

Toutes les prédictions effectuées sont sauvegardées dans le répertoire `gleam/predictions`. Le nom des fichiers suit la convention `année et pays d'entraînement _to_ année et pays de la prédiction.tif`. À titre d'exemple, la figure 6.6 représente la répartition de la population colombienne en 2017 telle que le modèle entraîné sur le Brésil en 2015 l'a prédite d'après l'image satellite de 2017. La figure 6.7 prédit la répartition de la population brésilienne en 2017. Le blanc correspond à 0 habitant et le noir à 500 ou plus.

La population totale prédite pour le Brésil en 2017 est 272478315 (environ 30% de plus que la réalité [12]), et 31904347 (environ 35% de moins que la réalité [12]) pour la Colombie. En multipliant les valeurs obtenues pour la Colombie par 1.5, on devrait obtenir une estimation de meilleure qualité, comme on a pu le constater en faisant une mise à l'échelle de la prédiction générée pour le Portugal. Ce n'est cependant pas une astuce sur laquelle on devrait dépendre, car on aimeraient pouvoir appliquer ce modèle à des pays pour lesquels on ne possède pas d'estimation fiable de la population totale.

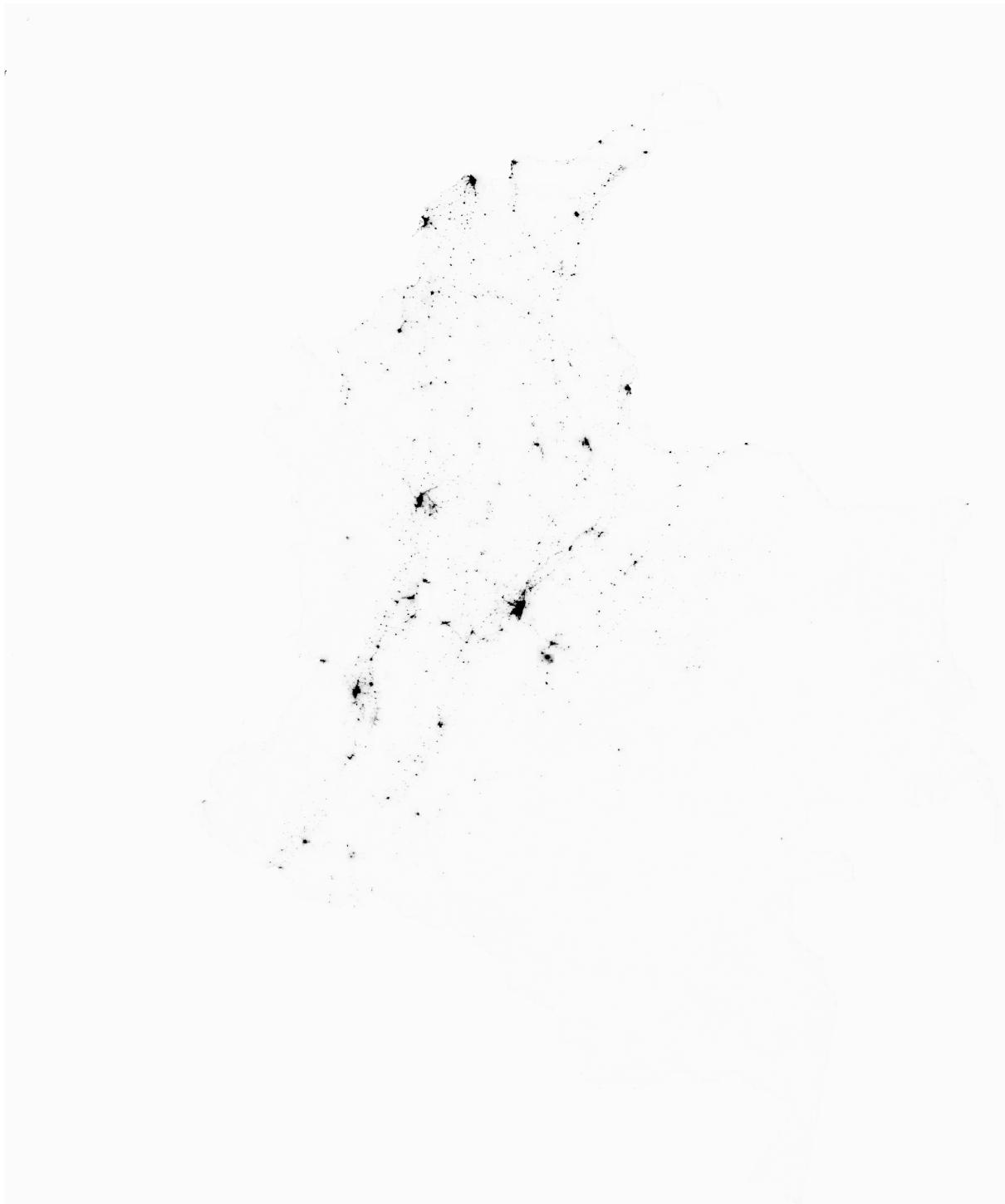


FIGURE 6.6 – Distribution de la population en Colombie en 2017 selon le modèle entraîné sur le Brésil en 2015.



FIGURE 6.7 – Distribution de la population au Brésil en 2017 selon le modèle entraîné sur le Brésil en 2015.

7 Conclusion

7.1 Discussion finale

Nous avons choisi de mettre en relation les images satellites nocturnes et la répartition géographique de la population. Cette décision a été justifiée par la disponibilité de grilles de population de haute résolution, une corrélation suffisante et l'utilité potentielle d'un modèle capable de générer rapidement une grille de population à jour.

Le modèle obtenu n'est pas parfait. Les sources erreurs sont cependant identifiées : données d'entraînement partiellement à jour, différences économiques entre pays et entre les quartiers d'une ville. Nous ne sommes pas parvenu à quantifier dans quelles proportions l'erreur est due au surapprentissage, aux données vieillissantes, au post-traitement de la prédiction ou aux différences entre pays.

La plus grosse difficulté rencontrée au cours de ce projet a été de travailler avec si peu de données de validation. Seuls une poignée de pays possèdent une grille de population exploitable, et les différences économiques ne permettent pas de valider le modèle sur un pays différent que celui qui a servi à l'entraîner. La seule réelle validation effectuée est la validation croisée sur l'image satellite du Brésil en 2015.

Malgré cela, les résultats obtenus sont prometteurs. Si l'on garde à l'esprit les limitations du modèle, telles que la répartition erronée à l'intérieur des villes ou les différences de l'utilisation de l'éclairage entre pays, on peut réaliser des estimations d'une précision géographique étonnante. Les prédictions manquent en fiabilité, mais peuvent se révéler très utiles lorsqu'elles sont faites sur les pays pour lesquels on ne dispose que de très peu de données démographiques. De plus, elles sont peu coûteuses à générer. Enfin, la résolution des cartes obtenues excède très largement celle des grilles de population issues d'études poussées pour la très grande majorité des pays, même en Europe.

7.2 Pour aller plus loin

Les possibilités ne s'arrêtent pas là. Avec plus de temps et de connaissances, en sortant du cadre fixé par le cahier des charges ou en attendant d'avoir plus de données à disposition, il serait possible de partir dans différentes directions :

- Améliorer l'algorithme de post-traitement qui distribue la population à l'intérieur d'une tuile prédictive de 256 km^2 .
- Identifier les variables qui permettraient de garantir que deux pays sont compatibles, et que l'un puisse servir de base pour faire une prédiction sur l'autre.

- Attendre la mise à jour de la grille de population globale ou utiliser des grilles spécialisées pour un pays en particulier, en supposant qu'elles soient plus fiables.
- Combiner les images satellites nocturnes avec celles de jour. La forme, le nombre et la taille des habitations donnent des informations supplémentaires précieuses, mais plus longues à traiter.
- Générer une image satellite nocturne ou une grille de population sur la base de celles des années précédentes afin d'obtenir les cartes des années à venir.
- Observer l'évolution des cartes mensuelles au cours d'une année "typique" afin d'obtenir des données saisonnières telles que l'impact des fêtes de fin d'année sur la lumière émise en Amérique du Nord.

Bibliographie

- [1] Jason BROWNLEE. *Crash Course in Convolutional Neural Networks for Machine Learning*. 24 juin 2016. URL : <https://machinelearningmastery.com/crash-course-convolutional-neural-networks/> (visité le 23/04/2018).
- [2] Jason BROWNLEE. *Evaluate the Performance Of Deep Learning Models in Keras*. 26 mai 2016. URL : <https://machinelearningmastery.com/evaluate-performance-deep-learning-models-keras/> (visité le 23/04/2018).
- [3] Jason BROWNLEE. *Handwritten Digit Recognition using Convolutional Neural Networks in Python with Keras*. 27 juin 2016. URL : <https://machinelearningmastery.com/handwritten-digit-recognition-using-convolutional-neural-networks-python-keras/> (visité le 23/04/2018).
- [4] Jason BROWNLEE. *Object Recognition with Convolutional Neural Networks in the Keras Deep Learning Library*. 1^{er} juil. 2016. URL : <https://machinelearningmastery.com/object-recognition-convolutional-neural-networks-keras-deep-learning-library/> (visité le 23/04/2018).
- [5] Jason BROWNLEE. *Regression Tutorial with the Keras Deep Learning Library in Python*. 9 juin 2016. URL : <https://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/> (visité le 23/04/2018).
- [6] CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK-CIESIN-COLUMBIA UNIVERSITY. *Gridded Population of the World, Version 4 (GPWv4) : Population Count Adjusted to Match 2015 Revision of UN WPP Country Totals, Revision 10*. 2017. DOI : [10.7927/h4jq0xzw](https://doi.org/10.7927/h4jq0xzw).
- [7] NOAA's National Geophysical Data CENTER et US Air Force Weather AGENCY. *Version 1 VIIRS Day/Night Band Nighttime Lights*. URL : https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html (visité le 15/07/2018).
- [8] NOAA's National Geophysical Data CENTER et US Air Force Weather AGENCY. *Version 4 DMSP-OLS Nighttime Lights Time Series*. URL : <https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html> (visité le 14/06/2018).
- [9] Patrick DOUPE et al. "Equitable development through deep learning". In : *Proceedings of the 7th Annual Symposium on Computing for Development - ACM DEV '16*. ACM Press, 2016. DOI : [10.1145/3001913.3001921](https://doi.org/10.1145/3001913.3001921). URL : <https://doi.org/10.1145/3001913.3001921>.
- [10] Christoph GOHLKE. *Unofficial Windows Binaries for Python Extension Packages*. URL : <https://www.lfd.uci.edu/~gohlke/pythonlibs/> (visité le 09/07/2018).
- [11] Natural Earth. *Admin 0 – Countries*. Version 4.0.0. Natural Earth. 21 mar. 2018. URL : <https://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-0-countries/> (visité le 07/06/2018).
- [12] World Bank Open Data. Version version. note. The World Bank. 12 oct. 2016. URL : <https://data.worldbank.org/> (visité le 07/06/2018).

- [13] *World Population Prospects 2017*. Version Révision de 2017. United Nations Department of Economic et Social Affairs, Population Division. 7 déc. 2017. URL : <https://esa.un.org/unpd/wpp/> (visité le 23/03/2018).
- [14] *Worldview*. NASA EOSDIS. URL : <https://worldview.earthdata.nasa.gov/> (visité le 06/06/2018).

8 Authentification

Par la présente, je soussigné, Antoine Friant, déclare avoir réalisé seul ce travail et ne pas avoir utilisé d'autres sources que celles citées dans la bibliographie.

Date

Signature

Nom complet

Table des figures

5.1	Outil de visualisation NASA Worldview [14].	4
5.2	Image satellite quotidienne servie par NASA Worldview [14], représentant la Grande-Bretagne et son climat nuageux.	5
5.3	Une tuile de l'image de 2016 montrant la ville de Dallas (USA) après avoir été mise en couleurs négatives.	5
5.4	Image globale annuelle (2016) reconstituée à partir de tuiles téléchargées, puis mise en couleurs négatives.	6
5.5	Extrait de la grille de population [6] rendu avec QGIS. Le blanc indique une absence d'habitants, le noir indique au moins 1000 habitants par kilomètre carré.	8
5.6	Quantité de lumière perçue depuis l'espace, population et PIB de la France entre 1992 et 2013.	10
5.7	Quantité de lumière perçue depuis l'espace, population et PIB de la Chine entre 1992 et 2013.	11
5.8	Quantité de lumière perçue depuis l'espace, population et PIB du Japon entre 1992 et 2013.	11
5.9	Quantité de lumière perçue depuis l'espace, population et PIB de l'Arménie entre 1992 et 2013.	12
5.10	Pays placés par population et luminosité totale émise sur une échelle logarithmique en 2013, colorés par indice économique.	13
5.11	Pays placés par PIB en USD et luminosité totale émise sur une échelle logarithmique en 2013, colorés par indice économique.	13
5.12	Pays placés par consommation en électricité en millions de kWh et luminosité totale émise sur une échelle logarithmique en 2013, colorés par indice économique.	14
5.13	Superposition de l'image satellite nocturne (bleu ciel) et de la grille de population (rose).	15
5.14	Effet de l'augmentation de la résolution de la grille de population lors de la fusion	16
5.15	Valeur de luminosité et de population mondiale pour chaque km ² (année 2000)	16

5.16 Valeur de luminosité et de population pour chaque 0.25km ² du Brésil (année 2015)	17
6.1 Topologie de la toute première version du réseau de neurones.	20
6.2 Moyennes des erreurs absolues sur 100 itérations.	20
6.3 Moyennes des erreurs au carré (fonction objectif) sur 100 itérations.	21
6.4 Grille de population pour la frontière nord entre l'Espagne et le Portugal illustrant les différences de qualité des données de population d'un pays à l'autre.	22
6.5 Résultat de la soustraction de la grille de population SEDAC [6] par la prédiction du modèle entraîné sur le Brésil en 2015. La ville représentée est Brasilia. Le blanc signifie une surestimation de la population de 500, le noir signifie une sous-estimation de la population de 500 par pixel (0.25 km ²).	25
6.6 Distribution de la population en Colombie en 2017 selon le modèle entraîné sur le Brésil en 2015.	29
6.7 Distribution de la population au Brésil en 2017 selon le modèle entraîné sur le Brésil en 2015.	30