UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE

des Sciences de la Terre et de l'Environnement (ED 413)

Laboratoire Image Ville Environnement (UMR CNRS 7362)

THÈSE présentée par :

Romain Wenger

soutenue le : **20 mars 2023**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Géographie – Télédétection/Géomatique

---

Apport des images Sentinel-1&2 et des méthodes d'apprentissage profond pour la cartographie et le suivi des modes d'occupation des sols

---

THÈSE dirigée par :
  Anne Puissant   Professeure, université de Strasbourg
  Germain Forestier   Professeur, université de Haute-Alsace

RAPPORTEURS :
  Clément Mallet   Directeur de recherche, université Gustave Eiffel
  Laurence Hubert-Moy   Professeure, université Rennes 2

AUTRES MEMBRES DU JURY :
  Charlotte Pelletier   Maître de conférences, université Bretagne Sud
  David Sheeren   Maître de conférences, Toulouse INP-ENSAT
  Jonathan Weber   Maître de conférences, université de Haute-Alsace
  Lhassane Idoumghar   Professeur, université de Haute-Alsace

Thèse de doctorat

Filière : Géographie

Spécialisation : Télédétection/Géomatique

Par

M.   Romain Wenger

Apport des images Sentinel-1&2 et des méthodes d'apprentissage profond pour la cartographie et le suivi des modes d'occupation des sols

Soutenance publique le 20 mars 2023

DR.	Clément Mallet	LASTIG - UGE	Rapporteur
Prof.	Laurence Hubert-Moy	LETG - Rennes-2	Rapporteure
MCF.	Charlotte Pelletier	IRISA - UBS	Examinatrice
MCF.	David Sheeren	DYNAFOR - INP-ENSAT	Examinateur
Prof.	Anne Puissant	LIVE - UNISTRA	Directrice de thèse
Prof.	Germain Forestier	IRIMAS - UHA	Co-directeur de thèse
MCF.	Jonathan Weber	IRIMAS - UHA	Encadrant
Prof.	Lhassane Idoumghar	IRIMAS - UHA	Encadrant

To my parents, to Gwendoline. . .

Acknowledgements

« *Wine is liquid geography.* » - Erik Orsenna, repris par Guillaume Piasny

Merci à l'équipe MSD de l'IRIMAS et merci à Cédric Wemmert pour son accompagnement depuis ma première année de post-bac en DUT informatique.

J'émets une attention particulière à mes amis de très longue date, David M., Clément F., Patrick V., Florian T., Mael M., Raphael A., David F. et Cécile I..

Je tiens également à remercier l'ensemble de ma famille et ma belle-famille, dont particulièrement mes parents qui m'ont épaulé tout au long de mon parcours quels que soient mes choix.

Enfin, merci Gwendoline pour ton soutien depuis près de 8 ans maintenant. Les derniers mois ont été rudes, les six à venir également, mais bientôt, nous regarderons toutes ces étapes de franchies avec un sourire radieux.
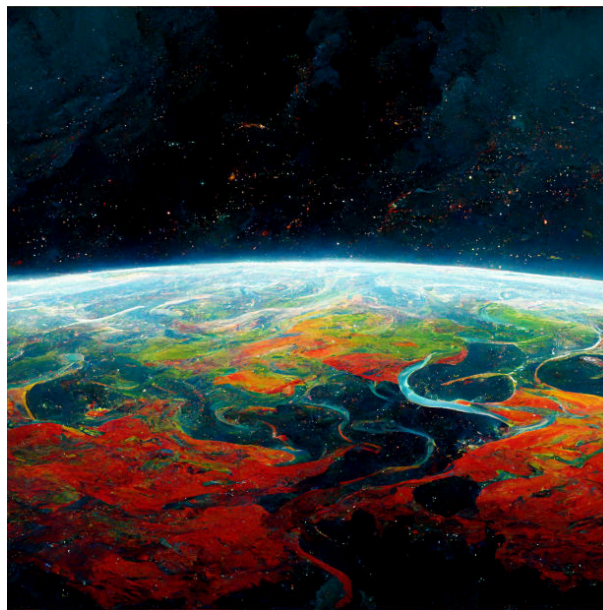
"Remote Sensing of Earth" by Midjourney (https://www.midjourney.com/) - Stable Diffusion network

Contents

List of Figures

List of Tables

Introduction

<div style="text-align: right; font-size: 3em; font-weight: bold;">1</div>

Over the past 150 years, humans have had a major impact on land cover, particularly through urban sprawl and industrialization, which has resulted in the release of substantial amounts of carbon into the atmosphere. This process has accelerated in recent decades and has become a major environmental concern worldwide [Lambin et al., 2003]. Land Use Land Cover (LULC) plays an important role in the exchange of greenhouse gases between the surface and the atmosphere. For example, deforestation releases large amounts of carbon into the atmosphere and affects cloud cover, evapotranspiration and surface albedo, which consequently affects climate change. Today, forests represent 30% of the land surface; however, 8,000 years ago, they represented 50% [Ball, 2001]. In contrast, afforestation and reforestation, although on smaller scales, can be positive for soils and ecosystems as they absorb carbon from the atmosphere. For example, in Europe, the total surface of artificial areas doubled between 1950 and 1990 [Gerard et al., 2010]. Changes in LULC can have both negative and positive impacts on humans but also have unexpected or unintended consequences [DeFries et Belward, 2000]. The Anthropocene era has seen a significant transformation of the Earth's surface, with more than half of it being altered by human activity, as reported by Hooke et al. [2013]. It is crucial to have accurate and up-to-date LULC maps to quantify and evaluate the impact of climate change on natural ecosystems. These maps also play a significant role in urban planning and management. Furthermore, they provide valuable information for scientists, as LULC has been recognized as an Essential Climate Variable (ECV) [Bontemps et al., 2015]; LULC, plays a critical role in land carbon models and can be both a cause and a consequence of climate change. It is at the heart of the work of the Intergovernmental Panel on Climate Change (IPCC), which uses it as support for many models that address carbon issues. It is therefore a key indicator that describes the evolution of the Earth's climate.

The terms "Land Use" and "Land Cover" have specific meanings and some fundamental differences in how they describe the land. They should not be used interchangeably. Following the INSPIRE European directive, land cover and land use have two distinct

definitions (INSPIRE - Infrastructure d'information géographique dans la Communauté européenne - 2007/2/EC) [Giri, 2012]:

- **Land Cover**: "*Physical and biological cover of the earth's surface including artificial surfaces, agricultural areas, forests, (semi) natural areas, wetlands, water bodies*". For Meyer et Turner [1992], land cover refers to the observed biotic and abiotic assemblage of the Earth's surface and immediate subsurface.

- **Land Use**: "*Territory characterized according to its current and future planned functional dimension or socioeconomic purpose (e.g., residential, industrial, commercial, agricultural, forestry, recreational)*".

Following these definitions, land cover refers to the surface or biophysical cover on the ground. Land use indicates how people are using the ground, such as for agriculture, airport zones or residential zones. These uses are difficult components to observe due to the differences between expected and actual use. We therefore see physical artifacts of use [Giri, 2012]. For example, forests seen through the prism of the land cover are generally a compact set of very dense trees. In contrast, forests can have multiple uses. They can be used for timber production, biodiversity conservation, hunting or fuel-wood production [Mercier et al., 2019]. Land cover is therefore rarely separated from land use to form a single definition.

1.1 Satellite imagery and land use/land cover mapping

For several decades, numerous Earth observation missions have been carried out with different objectives, including the visualization and monitoring of changes resulting from human activities. The first work on LULC mapping from remotely sensed data dates back to the mid-1940s. Francis J Marschner set out to map the whole United States by hand using aerial photography [Marschner, 1950]. The years following his work were dedicated to the development of space missions. It was not until the late 1960s that the scientific community began to take an interest in space LULC monitoring. The launch of the US Landsat-1 satellite in 1972 was a turning point for Earth observation. This sensor was the first to offer images for civilian uses. The images offered by this sensor were available in four bands, three in the visible and one in the near infrared at a spatial resolution of 80 m. Since then, several studies have been carried out using image processing techniques to produce LULC maps [Welch et al., 1975, Rogers et al., 1975, Todd, 1977]. Limited spatial resolution meant that typological diversity could not go beyond 5 to 8 classes: "Urban," "Agricultural," "Wooded," "Water," "Wetland," and "Bare Land" for Odenyo et Pettry [1977].

Between 1986 and 2002, in Europe, the CNES (Centre National des Etudes Spatiales) launched five SPOT (Satellite pour l'Observation de la Terre) Earth observation satellites, offering high spatial resolution imagery (6 m to 20 m for multispectral and 2.5 m to 10 m for panchromatic). Subsequently, SPOT Image and Atrium Satellites continued the programme by launching two new satellites in 2012 and 2014 with finer spatial resolution (1.5 m for panchromatic and 6 m for multispectral). Many works have taken advantage of the advent of this high spatial resolution to provide more accurate and detailed LULC mapping [Muchoney et Haack, 1994, Franklin et al., 2011].

At the end of the 1990s, NASA launched the Moderate-Resolution Imaging Spectro-radiometer (MODIS) from the Earth Observing System (EOS). The goal was to assess environmental dynamics (i.e., atmosphere, oceans, biosphere and land) at a large scale with a moderate spatial resolution. Two satellites were launched through the MODIS program, MODIS Terra and MODIS Aqua. These sensors have the advantage of providing a complete image of the Earth every 1-2 days and allow the detection of large-scale changes over short time intervals. These changes can be vegetation dynamics mapping [Fang et al., 2018], deforestation [Rahman et al., 2013] or land cover change in Arctic permafrost landscapes [Muster et al., 2015]. One of their major advances has been the world map generated through MODIS imagery, MODIS Land Cover [Friedl et al., 2002]. This product, produced annually between 2001 and 2020 by NASA, describes land cover in 17 classes. The initial objective of this map was to monitor the dynamics of natural areas on a global scale as well as the expansion of urbanized areas. These land cover data use are five different classification schemes through supervised classification trees.

The objective of the European project GMES (Global Monitoring for Environment and Security), also known as Copernicus, is to provide to the European community up-to-date Earth Observation data for territorial monitoring. The Copernicus program was set up by the European Union in 2014. It collects and distributes data regarding the condition of the Earth. The program has also launched earth observation satellites, such as Sentinel-2. In recent years, this sensor has been a breakthrough for LULC mapping, as it acquires images from the Earth's surface through a multispectral sensor (MSI) at high spatial (10 meters) and temporal resolutions. The innovation of this sensor lies in its constellation of two satellites, Sentinel-2A (launched in 2015) and Sentinel-2B (launched in 2017), which reduce the revisit time by half (5 days instead of 10). Through the diversity of its spectral bands, the objective of these satellites was to continue the acquisition of land surface imagery. Many works have used these images to map LULC to assess and quantify changes in urban [Papadomanolaki et al., 2019b] and/or natural areas [Pacheco-Pascagaza et al., 2022].

The Copernicus program also developed a large-scale LULC product, Corine Land Cover. Its aim is to provide a LULC map at the European scale for several years, or milesims. The semantic class typology was selected according to the INSPIRE directive, which defines the relevant classes on which LULC mapping should focus. Corine land cover (Figure 1.1 (a)) is produced manually from photointerpretation of satellite imagery at 20- and 25-meter spatial resolutions and from images with higher spatial resolution for the densest areas (e.g., urban areas) at a scale of 1:100 000. This map is produced for all European Union countries, including some neighboring countries (e.g., Turkey or Serbia); 39 countries are covered by Corine Land Cover with 5 millesims to date: 1990-2000-2006-2012-2018. Corine land cover describes LULC in 3 levels of semantic classes. The minimum mapping unit (MMU) corresponds to 25 hectares ($\sim$ 35 football fields), which makes its use possible at a scale of 1:100,000. At this spatial scale, urban areas cannot be studied in detail. Indeed, roads and large-scale networks, whose MMUs are well below 25 Ha, are not represented, leading to requests for other LULC maps. To overcome this limitation, the Copernicus program has set up a product developed only over the largest cities in Europe (urban areas with more than 50,000 inhabitants), Urban Atlas (Figure 1.1 (b)), which allows the precise mapping of urban areas in 27 classes for the most detailed level. The minimum MMU is 0.25 ha, which is 100 times smaller than the Corine land cover product. Currently, Urban Atlas is available for 3 dates: 2006, 2012 and 2018.

In France, the Institut national de l'information géographique et forestière (IGN) has produced a national land use land cover map since 2013 that also meets the INSPIRE guidelines: the Occupation du Sol à Grande Échelle (OCS GE). The OCS GE is also organized into several levels of semantic classes and has 24 classes within its finest level. To produce this cartography of the territory, the IGN used different databases (BD TOPO, BD Foret or Registre Parcellaire Graphique, also known as RPG). The MMU for the OCS GE is 200 m2 for built-up artificial areas (e.g., dense built-up areas, residential areas), 500 m2 for nonbuilt-up artificial areas (e.g., large scale networks) and 2500 m2 for nonartificial areas (e.g., forests, water surfaces). Since 2022, within the framework of the land artificialization observatory, the IGN has relied on artificial intelligence (AI) to automate the initial production and update the production of geographical data describing LULC according to the OCS GE nomenclature (in 13 classes). These prototype data will be progressively disseminated. In parallel, several administrative regions are piloting the production of LULC maps adapted to their needs. For instance, the Alsace Region has started the production of a LULC map at a large scale through the Coopération pour l'Information Géographique en Alsace (CIGAL, started in 2008). Since the creation of the Grand-Est region, a large-scale

(1:10000) topographic database 'BD OCS GE2 - Base de Données Occupation des Sols à Grande Échelle du Grand-Est' has been produced for 2018/2019 (Figure 1.1 (c)). These regional databases are most often built from visual interpretation of aerial photographs and are time- and cost-consuming for the administration. In the Grand-Est region, the MMU is one of the finest LULC products presented to date, with 50 m2 for buildings and urban areas. The objective of this project, founded partially by the European Union, is to offer a better knowledge of the territory through a very fine mapping of urban areas. This product has four main semantic levels with a 5th class for urban areas that describe the degree of imperviousness of level 4 artificial classes (Imperviousness Built-Up, Imperviousness Not Built-Up, Pervious). At the finest scale, urban areas are described in 87 thematic classes.
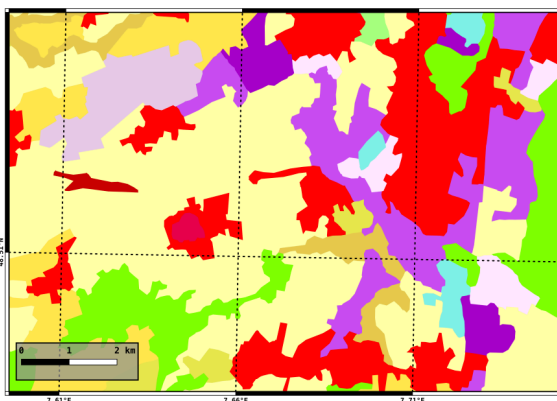
(a) *Corine Land Cover* 2012 over the south west of Strasbourg.
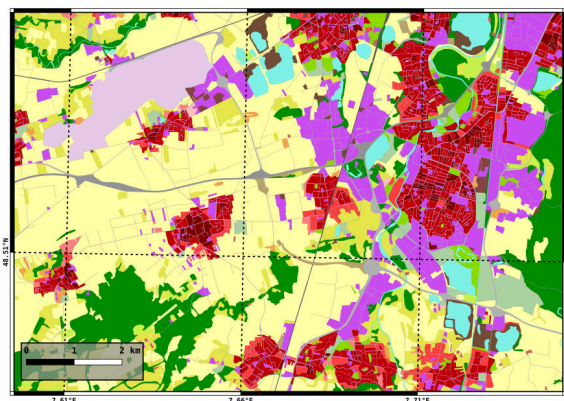
(b) *Urban Atlas* 2018 over the south west of Strasbourg.
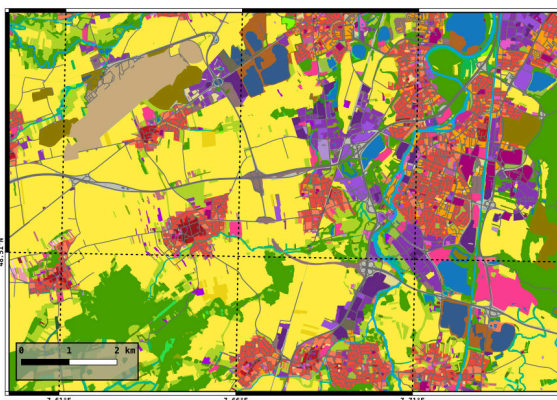
(c) *Occupation du Sol à Grande Échelle du Grand-Est 2019* over the south west of Strasbourg.
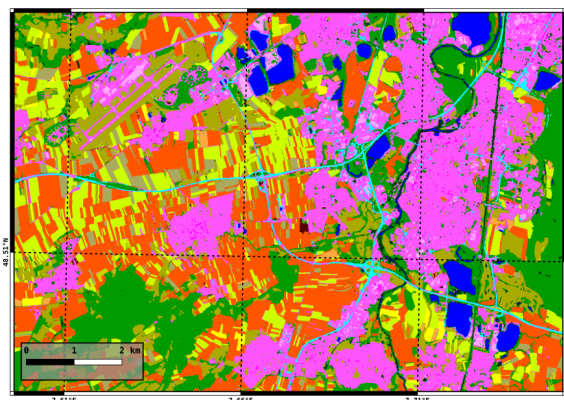
(d) *Occupation du SOl (OSO)* 2021 over the south west of Strasbourg.

Figure 1.1 – *Comparison between three Land Use Land Cover database over Strasbourg (Legends are in appendix A.1).*

Although accurate, visual interpretation is subject to many problems. First, the production time can be several months or years for most LULC data. For example, the Corine Land Cover product was only released to the community in 2015, almost three years later.

This delay raises issues in monitoring rapid phenomena such as crops or the evolution of natural areas. Additionally, production operators are limited in their interpretation of images and often do not take into account the multitemporal dimension of the images. However, this analysis of radiometric characteristics is important to identify natural landscapes, particularly agricultural landscapes, which have strong intratemporal variation. A bias can also be found in some classifications due to the perception of the territory, which can cause spatial heterogeneity in the mapped products between regions.

Artificial intelligence methods, particularly machine learning, have seen a renewed interest in recent years, thanks in particular to improvements in computing power. This term covers two main concepts: supervised and unsupervised learning. The first one requires an annotated dataset (or training set). Remote sensing and LULC mapping consist of images, pixels or groups of pixels depending on the application. In the remote sensing domain, the training data are most often built from existing LULC datasets considered as a reference. This step is the most important and requires precise knowledge of the study area to choose the appropriate data in accordance with the initial objective. This requirement can cause several problems, as these data may (1) not be consistent with the spatial resolution of the images used and/or (2) have a temporal shift in mismatch with the classes with high temporal variability. With the high volume of satellite imagery collected each day by multiple sensors, deep learning techniques based on neural networks have improved the results of so-called "classic" machine learning approaches.

Since 2015, with the open access Sentinel constellation, several LU and/or LC products have been created with national or worldwide coverage. In France, through the national data and services centercenter THEIA for continental surfaces (attached to the National Infrastructure Data Terra), several research laboratories have developed open-source processing chains based on Sentinel1&2 imagery. Since 2016, they have produced an open-access LULC map (named OSO) on the metropolitan national territory annually (millesim) at 10-meter spatial resolution using either machine learning approach [Inglada et al., 2017]. The typology evolves each year to provide users with a refined product with a greater variety in semantic classes. Since 2018, the production has been carried out by CNES MUSCATE (Figure 1.1 (d)), and laboratories continue to improve large-scale land cover mapping methods and to integrate them into the iota2 processing chain (e.g., superpixel classification).

Since 2020, three global LULC products have been opened to the community: Google's Dynamic World [Brown et al., 2022] (Figure 1.2 (c)), ESA's World Cover 2020 [Zanaga et al., 2021] (Figure 1.2 (a)) and Esri's 2020 Land Cover [Karra et al., 2021] (Figure 1.2 (b)). The objective for these three products is to offer multitemporal cartography, where an update

is performed every year for ESA's World Cover and Esri's 2020 Land Cover and in near real time for Google's Dynamic World at each acquisition of new satellite images. These products are generated automatically thanks to the progress of AI methods. Both Google's Dynamic World and Esri's 2020 Land Cover propose maps with 9 LULC classes whereas ESA's World Cover 2020 has added 2 more classes to obtain a typology in 11 classes. The ESA World Cover 2020 product uses classical machine learning methods (Random Forest) trained on hand-labeled pixels from a 100 × 100 m grid over 141,000 locations world-wide. The input satellite data are mainly from the optical Sentinel-2 (L2A) sensor and SAR Sentinel-1 (GRD) sensors. In addition, many other data (positional features, meteorological features or features from the DEMs) have been introduced into their models. The land cover product, Esri's 2020 Land Cover, uses deep learning methods trained on over 5 billion human-labeled Sentinel-2 pixels from 24,000 individual image tiles sized 510 x 510 pixels each.
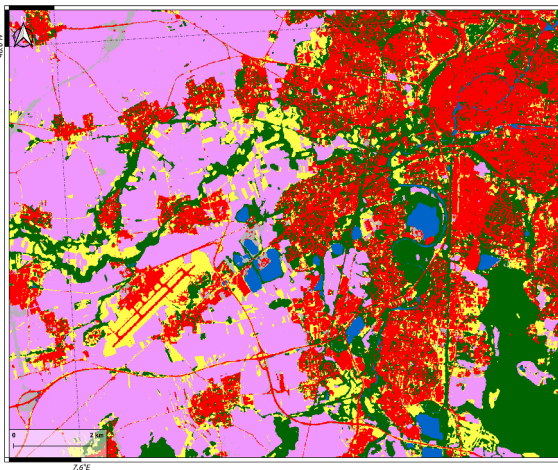
(a) *ESA World Cover* 2020 over the south west of Strasbourg.
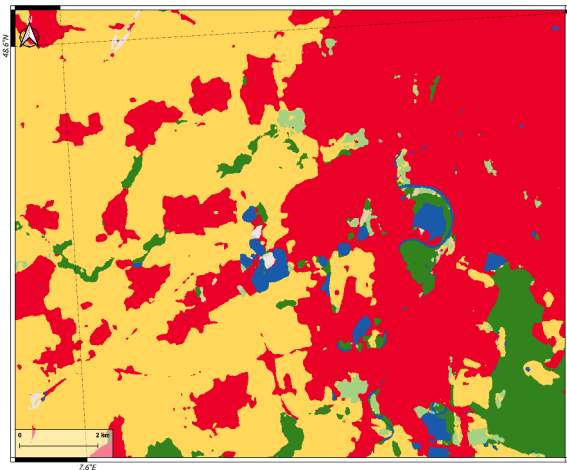
(b) *ESRI 2020 Land Cover* over the south west of Strasbourg.
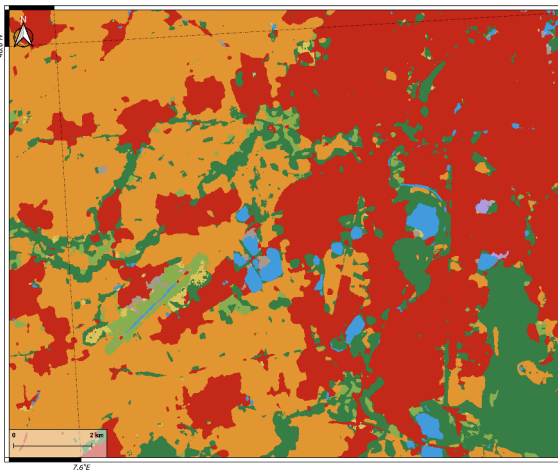
(c) *Google Dynamic World* from 01/01/2022 to 31/12/2022 over the south west of Strasbourg.
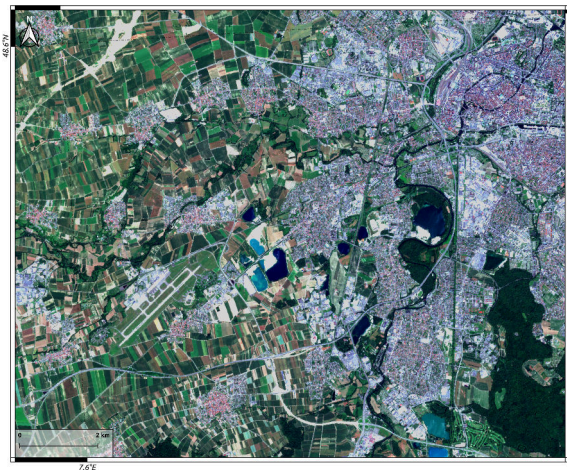
(d) *Sentinel-2 L2A* over the south west of Strasbourg.

Figure 1.2 – *Comparison between ESA World Cover, ESRI 2020 Land Cover and Google Dynamic World over Strasbourg. Legends are in appendices A.2, A.3 and A.4.*

Among this panel of LULC maps, several intrinsic characteristics can help end users understand and choose a mapping product adapted to their needs. They can be summarized in 4 criteria described in Table 1.1:

- **The spatial extent**: this describes the spatial cover of the maps. A map can be produced for different areas (specific, country, pan-European, world, etc.)

- **Date of production**: every map has a date of production and a date of distribution. Most vector LULC maps are distributed several months or years after the beginning of production. This information is important because maps with different production dates enable users to extract temporal dynamics.

- **The number of thematic classes** defines the semantic typology of an LULC product, which can be organized into a nomenclature describing different levels of classes from a general to a fine description of LULC. To obtain homogeneity in the nomenclature of products, the Food and Agriculture Organization (FAO) has set up a hierarchical land cover classification system (LCCS) that provides different levels of information according to use. In general, products contain 3 to 4 levels of classification, from the least detailed (often binary) to the most detailed.

- **Spatial resolution or cartographic scale**: Land use/land cover maps are also described by several geographical characteristics, such as the Minimal Mapping (MMU) for a vector map or the pixel size for a raster map. Any object smaller than this value will not be displayed on the map.

In conclusion, there are multiple LULC databases at different scales and formats, sometimes produced by digitization of aerial photographs and/or satellite images and other times by automatic processing, often using AI methods (Tableau 1.1). The accuracy and diversity of the semantic classes is often the result of a manual approach that is very time-consuming and involves a product that is out of date at the time of its release. In contrast, LULC products derived from AI approaches offer lower accuracy and a smaller number of thematic classes but are available and up to date at much shorter time intervals than manual products. This state-of-the-art LULC product is not necessarily exhaustive because each nation works on the development of a nationwide or worldwide product (e.g., Globeland30[1] which is a global geo-information public product provided by China to the United Nations [Fonte et al., 2017]) but it shows the diversity of existing and ongoing research works.

---

[1]http://www.globallandcover.com/Page/EN_sysFrame/dataIntroduce.html?columnID=81&head=product&para=product&type=data

Product name	Format	Resolution or Cartographic Scale	Production method	Classification Level	Number of semantic classes	Dates 'millesims'	Spatial Extent
Dynamic World	Raster	10 meters	Classification	1	9	Near real time	World
ESRI's 2020 Land Cover	Raster	10 meters	Classification	1	9	Each year since 2020	World
ESA's World Cover	Raster	10 meters	Classification	1	11	Each year since 2020	World
Corine Land Cover	Vector	1/100,000	Manual	3	44 at level 3	1990 - 2000 - 2006 - 2012 - 2018	Pan-European
Urban Atlas	Vector	1/10,000	Manual	1	27	2006 - 2012 - 2018	Pan-European
OSO	Raster and Vector	10 meters	Classification	1	17 to 23	Each year since 2016	France
OCSGE2	Vector	1/2,000	Manual	5 for urban areas and 4 for natural areas	53 for level 4	2019 - 2020	French Region

Table 1.1 – *Summary of LULC products*

1.2 Urban thematic classes in the LULC maps

For countries with rapid growth where reference data are often missing, products resulting from automated processes (e.g., methods based on artificial intelligence) classify the urban footprint into one class (mostly known as artificial surfaces). This situation is the case for the three recent products, Google's Dynamic World, ESA's World Cover 2020 and Esri's 2020 Land Cover, each of which provides a single "Built-Up Area" class. For several years, many "urban" products have been developed with the aim of monitoring artificial or impervious surfaces (Table 1.2). Each product extracts different information often in a raster format.

The Global Human Settlement Layer (GHSL) is a LULC product that describes the human settlement footprint on Earth. Developed by the JRC (Joint Research Center), several 'millesims' have been produced (1975, 1990, 2000, 2014) from Landsat images (30 m spatial resolution) by applying a supervised classification system to them. A built-up map with a spatial resolution of 38 meters (using Landsat images) is thus created, containing pixel information representing the average urban proportion on it. Since 2016, the JRC has offered the same mapping developed from Sentinel-1 images from the Copernicus program available at 20-meter spatial resolution (Figure 1.3 (b)).

Another product, the Global Urban Footprint (GUF), offers a finer global mapping (12 meters spatial resolution) of land cover in three classes: urbanized areas, nonurbanized areas and areas with no data coverage (e.g., oceans). TerraSAR-X and TanDEM-X are the two sensors used to produce this map. They are both synthetic aperture radars with a resolution of 3 meters. A classification is performed based on the speckle divergence to highlight the areas subject to backscattering [Esch et al., 2013]. Thus, not only are buildings taken into account, but all pixels with a high texture value are classified as buildings. In 2019, a global version of this product, named the World Urban Footprint, was released (https://urban-tep.eu/). This map, which is available at a resolution of 12 meters, was also produced from the 2015 satellite imagery (Figure 1.3 (a)).

The High Resolution Layer Imperviousness (*HRL Imperviousness*) is produced by Copernicus and offers two distinct maps: the first one translates the percentage of soil imperviousness, and the second one shows the evolution (the difference) between two millesims. The percentage of soil impermeability is calculated by a processing chain of multiseason images and uses the NDVI (Normalized Difference Vegetation Index) for a better distinction of impermeable surfaces. Beforehand, a stratification of the territory into two classes (built-up and nonbuilt-up) is performed. The classification is produced accord-

ing to a pixel approach, and the validation samples are extracted from the Corine Land Cover layer (Figure 1.3 (c)).
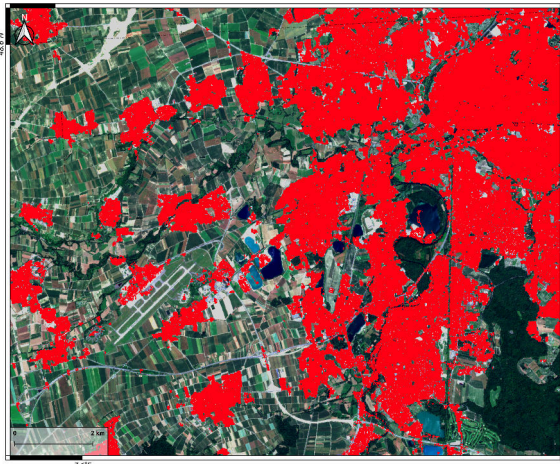
(a) *GUF* over the south west of Strasbourg (Urbanized Areas in red).
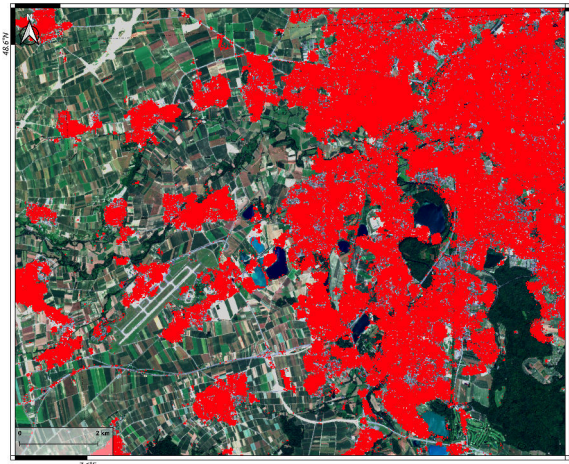
(b) *GHSL* over the south west of Strasbourg (Human Settlement Footprint in red).
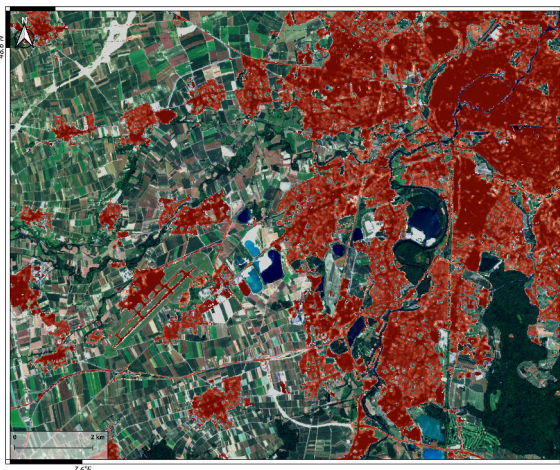
(c) *HRL Imperviousness* over the south west of Strasbourg (Imperviousness degree in red graduated color, from 0 to 100%).
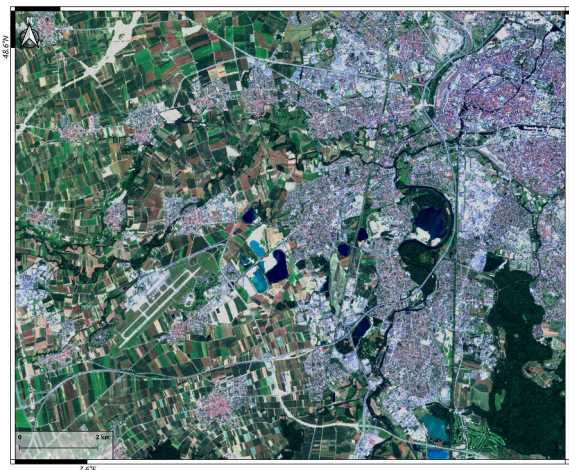
(d) *Sentinel-2 L2A* over the south west of Strasbourg.

Figure 1.3 – *Comparison between GUF, GHSL and HRL Imperviousness over Strasbourg.*

When users are interested in an urban domain application or focused on cities, existing LULC maps often poorly describe urban forms (morphology), also called urban fabrics (UF), defined as the specific spatial organization of basic components of the city. These components comprise the inner characteristics of the built-up environment.

To our knowledge, only OSO (which is based on Sentinel-2 imagery) proposes four semantic classes to describe UF (Figure 1.4). These classes are dense urban, sparse urban, industrial and commercial areas and roads. Based on so-called "pixel" approaches, the results contain many isolated pixels, causing a strong salt-and-pepper effect. They are

often confused with many other classes at high spatial resolution [Inglada et al., 2017]. In recent works, this mapping has been improved with the introduction of superpixel-based approaches [Derksen et al., 2019].

(a) *OSO* 2018 over Strasbourg.

(b) *Sentinel-2 L2A* over Strasbourg.
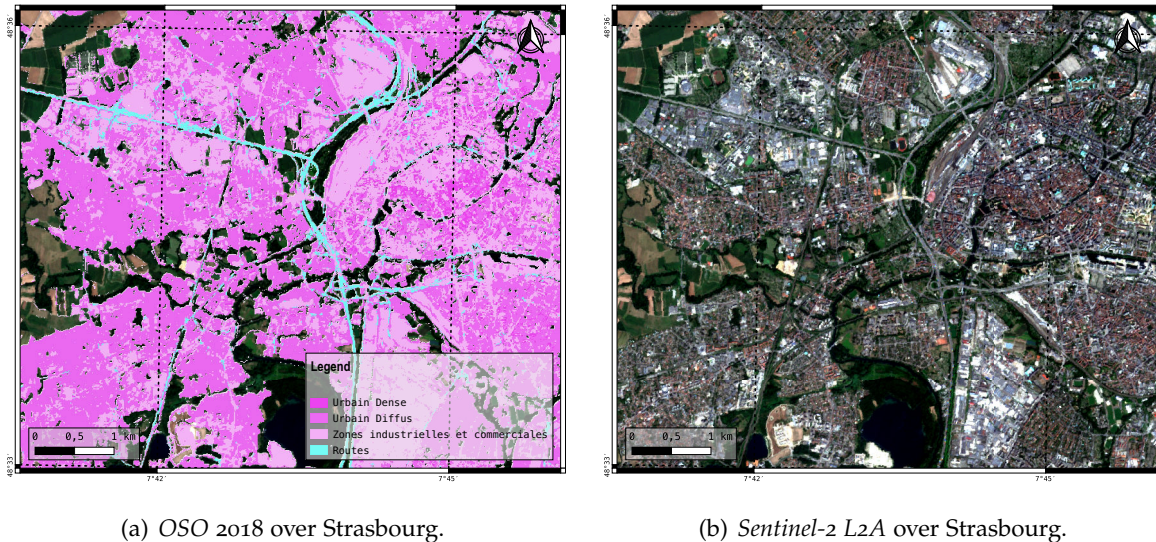
Figure 1.4 – *Urban Fabric classes for OSO over Strasbourg.*

In this context, very high spatial resolution (VHSR) imagery (defined here with a metric or submetric spatial resolution) is preferred to produce LULC maps. Indeed, these images allow a fine mapping of urban elements and objects. For example, Puissant et al. [2014] and Rougier et al. [2016] detected trees in urban areas from VHSR images (Pleiade and Quickbird) using machine learning methods (Random Forest). Other works tend to detect building footprints while omitting the urban fabric (UF) aspect [Zhang et al., 2018]. These machine learning methods are often based on an object-based approach. It is more complex to effectively detect UF [Merlin et al., 1988], which have a particular spatial organization, with methods that are difficult to reproduce and generalize [Huang et al., 2018, Postadjian et al., 2017]. Thus, building a fine scale urban map can be performed according to two strategies: (1) extracting one class from VHRS images (e.g., vegetation) with per-pixel or object-based approaches [Rougier et al., 2016, Li et al., 2016] and its combination with existing topographic datasets or (2) directly classifying urban objects or elements into several semantic classes of land cover (e.g., building, water, vegetation, bare soil) with per-pixel or object-based approaches [Ban et al., 2010, Postadjian et al., 2017].

These LC maps at VHRS are relevant inputs to build Local Climate Zone (LCZ) maps that are classically useful for urban climate models [Leconte et al., 2015]. The advantage of these maps is that they provide rich information on the spatial organization of the city where vegetation has an important role in the overall typology. Moreover, in the existing vector topographic databases such as 'BD OCS GE2', vegetation is categorized by use (e.g.,

urban park labeled as specialized vegetation class) and is explicitly separated from the dense, sparse and other specialized urban classes (for industrial or commercial activities). This class addresses an important need of end-users, as it answers many ecological and climatic issues (urban cool islands and many other ecosystem services) [Mexia et al., 2018]. However, VHSR images remain costly and are often unavailable as time series, and the models trained on these images are difficult to generalize due to the complexity of urban areas at this resolution [Momeni et al., 2016]. This process is also very time-consuming, as retraining of the models is necessary for most cities.

In this context, research on UF mapping based on HSR imagery is challenging. The temporal diversity of sensors and the diversity of acquisition modes (e.g., optical and SAR imagery) as well as the advent of new computational methods (deep learning and graphics card computing) have opened new opportunities for the classification of UF into several classes to improve LULC maps.

Currently, the only work that attempts to improve the mapping of UF (or patterns) is one developed by [Schiavina et al., 2022]. The authors propose a multiclass urban product named global building morphology (GHS-BUILT-C), which will be an extension of the GHSL product. However, this product is not yet available and and their pixel approach may suggest a salt and pepper effect in the result.

Product name	Format	Resolution or Cartographic Scale	Source Data	Number of semantic classes	Dates 'millesims'	Spatial Extent
GHSL	Raster	38m, 250m, 1km	Landsat imagery	2	1975, 1990, 2000, 2014, 2016, 2018	World
GUF (or WUF)	Raster	12m, 84m	TerraSAR-X, TanDEM-X	2	2012 and 2015	World
HRL Imperviousness	Raster	20m	RessourceSAT-2, SPOT-4/5, IRS-P6	2 and imperviousness percentage	2006, 2009, 2012, 2016	Europe
OSO	Raster and Vector	10m, 20m	Sentinel-2 and/or Landsat	17 to 23 and 4 for UF	2016, 2017, 2018, 2019, 2020, 2021	Europe

Table 1.2 – *Summary of UF products*

1.3 OUTLINE OF THE THESIS

1.3.1 Research questions and hypotheses

As described in the previous sections, accurate and timely LULC maps are important for a variety of applications, such as urban and regional planning and management, environmental monitoring, disasters and risk analysis. They may help tackle many significant large-scale challenges, such as increasing urbanization, global warming or the accelerating loss of biodiversity. Therefore, it is important to produce accurate LULC maps, especially by improving the mapping of urban areas through their characterization into several UF.

Therefore, the general problem of the thesis concerns the mapping of LULC and specifically their urban components in the era of big data, which provides a substantial amount of freely available geospatial data. The volume of collected or archived satellite imagery has been increasing from terabytes to petabytes and even to exabytes in recent years [Zhang et Li, 2022]. While new opportunities are provided by AI technologies and massive data for LULC mapping, challenges also emerge for LULC and UF mapping, which has strong climatic issues, from using these high spatial, temporal and multimodal satellite imagery (e.g., Sentinel1&2). Moreover, it is still challenging to integrate multisource remotely sensed data.

Their high temporal resolution combined with their high spatial resolution make these sensors major assets in inventories and in the detection of temporal dynamics, whether at short times (e.g., landslides, earthquakes) or long times (e.g., urban sprawl, sea level increases). Moreover, these sensors have opened new perspectives for LULC by adding high temporal dimensions uncommon to most existing sensors. SAR imagery, through Sentinel-1, and optical imagery, through Sentinel-2 and its spectral diversity, provide important new properties for LULC mapping.

The evolution of computing power and massive data management has allowed the development of new methods based on machine learning and deep learning. One of machine learning's advantages is its ability to handle high-dimensional input and map LULC classes with complex properties. With the increasing amounts and types of remote sensing imagery and geospatial big data available, as well as the availability of cost-effective and powerful computing tools and storage solutions, machine learning has become more widely used for analyzing larger and more complicated datasets, resulting in more accurate LULC mapping at larger scales.

In this context, the main objective of this thesis is to **show the potential of massive high spatial resolution imagery (S1&S2) for mapping land use land cover at large-scale**

and analyzing its dynamics. The general hypothesis is that multimodal and multitemporal high spatial resolution imagery can improve the classification of LULC, especially for UF. Several methodological subobjectives can be derived:

- (1) evaluate the performance of classification methods based on *deep learning* approaches for UF mapping,

- (2) evaluate the synergy of multitemporal and multimodal satellite imagery provided by the Sentinel- 1&2 satellite constellation for UF and LULC mapping and

- (3) evaluate the spatial and temporal inference capacity of the proposed classification models to generalize the proposed approach for mapping large-scale territories and analyzing their dynamics.

1.3.2 Thesis organization

This document is organized into six chapters including the introduction section and the state-of-the art chapter (Chapter 2) and according to the papers that have been written during the corresponding PhD work: 1 in a double-blind peer-reviewed congress international journal (Chapter 3), and 2 published papers in peer-reviewed international journals, related to the first (Chapter 4) and to the second objectives of this thesis (Chapter 5). Another paper concerning the third objective (Chapter 6) corresponds to a short paper accepted in an international conference focused on urban areas (JURSE 2023).

- The second chapter focuses on the evolution of artificial intelligence methods and their use for LULC. As deep learning is the main subject of this thesis, machine learning methods will only be an introduction to the chapter with the presentation of existing works and the limits of these methods compared to deep learning approaches.

- The third chapter describes the **study area and the creation of a multimodal and multitemporal land use land cover dataset: the MultiSenGE dataset**. The objective is to prepare an LULC reference dataset over the study area [Wenger et al., 2022a]. In fact, this objective represents a major challenge in the remote sensing field, as no such dataset was available to the community.

- The fourth chapter aims to evaluate the potential of **monotemporal Sentinel-2 imagery and semantic segmentation approaches** to improve the UF classes into LULC maps through feature fusion methods [Wenger et al., 2022c].

- The fifth chapter addresses **the combination of multitemporal and multimodal imagery using the MultiSenGE dataset** [Wenger et al., 2023b]. The objective is to use feature fusion methods that take into account multitemporal and multimodal imagery and combine them to perform LULC mapping and to evaluate the contribution of these features compared to monotemporal optical imagery.

- The last chapter (6) aims to **investigate spatial and temporal inference** to explore the reproducibility of the methods developed in the previous chapter through two global aspects: spatial inference from the perspective of large-scale LULC production and temporal inference to detect changes (e.g., urban sprawl with construction sites) and to analyze LULC dynamics.

During this thesis, all the methods have been developed using open source softwares and libraries. In addition, optical (Sentinel-2) and SAR (Sentinel-1) imagery were also freely downloaded via different hubs (i.e. **Theia**[2] hub and **Scihub Copernicus**[3]). All the *Deep and Machine Learning* applications were performed using free packages available through **Python**[4] programming language, whether **Keras**[5] for Deep Learning applications or **Scikit-Learn**[6] for Classical Machine Learning applications. To process Sentinel-2 satellite images and also reference data, **GDAL**[7] and **RasterIO**[8] were mainly used. The downloading and pre-processing of Sentinel-1 images was done with **s1tiling**[9] developed by the CNES (Centre National d'Etudes Spatiales). Finally, the visualization and mapping productions were made with **QGIS**[10] software.

*This thesis has been founded by Agence Nationale de la Recherche (ANR) through ANR Exploitation de masses de données hétérogènes à haute fréquence temporelle pour l'analyse des changements environnementaux [**TIMES**, ANR-17-CE23-0015]. The objectives of the TIMES project is to **produce new knowledge on the dynamics landscape objects from the massive exploitation of this big geospatial data with the objective to develop and validate novel data processing and analysis methods for environmental monitoring of landscape objects**. The proposed methods will be able to tackle highly heterogeneous datasets (point cloud data, aerial and satellite images) analyzed at very high temporal frequency. Also, this thesis was supported by the French*

---

[2] https://www.theia-land.fr/
[3] https://scihub.copernicus.eu/
[4] https://www.python.org/
[5] https://keras.io/
[6] https://scikit-learn.org/
[7] https://gdal.org/
[8] https://rasterio.readthedocs.io/
[9] https://gitlab.orfeo-toolbox.org/s1-tiling/s1tiling
[10] https://www.qgis.org/

1.3.3 Scientific contributions

Research papers in peer-reviewed journals:

- **Wenger, R.**, Puissant, A., Weber, J., Idoumghar, L., & Forestier, G. (2022a). Multi-SenGE: a multimodal and multitemporal benchmark dataset for land use/land cover remote sensing applications. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3–2022, 635–640. https://doi.org/10.5194/isprs-annals-V-3-2022-635-2022

- **Wenger, R.**, Puissant, A., Weber, J., Idoumghar, L., & Forestier, G. (2022c). U-Net feature fusion for multi-class semantic segmentation of urban fabrics from Sentinel-2 imagery: an application on Grand Est Region, France. *International Journal of Remote Sensing*, 43(6), 1983–2011. https://doi.org/10.1080/01431161.2022.2054295

- **Wenger, R.**, Puissant, A., Weber, J., Idoumghar, L., & Forestier, G. (2023b). Multi-modal and multitemporal land use/land cover semantic segmentation on Sentinel-1 and Sentinel-2 imagery : an application on MultiSenGE dataset. *Remote Sensing*; 15(1):151. https://doi.org/10.3390/rs15010151

Communications in international conferences:

- **Wenger R.**, Puissant A., Weber J., Idoumghar I., & Forestier G., 2022, Multimodal and multitemporal semantic segmentation and scene classification dataset for remote sensing applications, *Living Planet Symposium*, 23-27 May 2022, Bonn, Germany. (poster presentation)

- **Wenger R.**, Puissant A., Weber J., Idoumghar I., & Forestier G., 2022, MultiSenGE: a multimodal and multitemporal benchmark dataset for land use/land cover remote sensing applications., *XXIV ISPRS Congress*, 6-10 June 2022, Nice, France. (oral presentation)

- **Wenger R.**, Puissant A., Weber J., Idoumghar I., & Forestier G., 2023, Exploring inference of a land use and land cover model trained on MultiSenGE dataset, Joint Urban Remote Sensing Event, 17-19 May 2023, Heraklion Crete, Greece. (oral presentation)

---

[11]https://hpc.pages.unistra.fr/

Also, an innovative multitemporal and multimodal Land Use Land Cover dataset[12] has been produced covering the entire Grand-Est region:

- **Wenger, R.**, Puissant, A., Weber, J., Idoumghar, L., & Forestier, G. (2022b). A new remote sensing benchmark dataset for machine learning applications: MultiSenGE. *Zenodo*. https://doi.org/10.5281/zenodo.6375466

This thesis work follows on from the master's work on urban areas. Thus, a scientific communication as well as a paper have been presented and are being reviewed in an international peer-reviewed journal:

- **Wenger R.**, Michea D., & Puissant A., 2021, Automated Urban Footprint Mapping Over Large Areas, EARSeL, 30 March - 1 April 2021, Liège, Belgium (Joint Virtual Workshop). (oral presentation)

- **Wenger R.**, Puissant A. & Michea D., 2023a, Towards an annual Urban Settlement map in France at 10m spatial resolution using a method for massive streams of Sentinel-2. *Submit, unpublished.*

---

[12]https://multisenge.github.io/

FROM MACHINE LEARNING TO DEEP LEARNING IN LAND USE LAND COVER MAPPING

2

CONTENTS

With the increasing need for up-to-date LULC maps, scientific community has started to use machine learning methods for several years. This chapter presents the evolution of machine learning methods for knowledge extraction and classification of LULC. This chapter focuses on deep learning and details the methods used to combine multitemporal and multimodal remote sensing imagery.

2.1 Machine learning versus deep learning

Artificial Intelligence is the global ensemble containing Machine Learning and Deep Learning (Figure 2.1). Machine learning is a subfield of artificial intelligence that involves the development of algorithms and statistical models that enable computers to learn from and make predictions or decisions without being explicitly programmed to perform a specific task. Machine learning is based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.
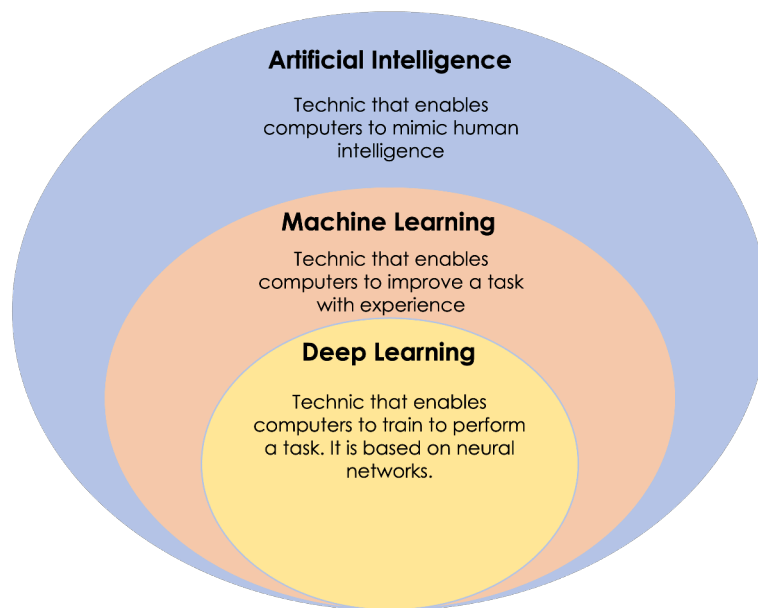
Figure 2.1 – *Artificial Intelligence versus Machine Learning versus Deep Learning*

Machine learning and deep learning have gained a lot of popularity in recent years. Both techniques involve the use of algorithms to analyze data and make predictions or decisions.

Machine learning algorithms use a set of labeled training data to learn a mathematical model that can be used to make predictions or decisions on new, unseen data. The goal of machine learning is to find the best possible model that can generalize to new data, and it relies on a number of different techniques, such as decision trees, Support Vector Machines, and random forests.

On the other hand, deep learning algorithms use a large number of layers of interconnected neurons to learn complex representations of data. These algorithms are able to learn highly non-linear relationships in data, and they are particularly well-suited for tasks such as image and speech recognition. Deep learning algorithms are typically trained using large amounts of data and powerful computing resources, such as graphics processing units (GPUs).

Overall, machine learning and deep learning are both powerful tools for solving com-

plex problems in artificial intelligence, and they have been applied to a wide range of applications in areas such as natural language processing, computer vision, and robotics. In general, machine and deep learning approaches are twofold: unsupervised learning and supervised learning. The main difference between the two is that supervised learning uses labeled training data, while unsupervised learning uses unlabeled data.

According to Singh et al. [2016], "supervised machine learning is the construction of algorithms that are able to produce general patterns and hypotheses by using externally supplied instances to predict the fate of future instances.". In supervised learning, the training data is labeled to indicate the correct input for each example; in a classification problem, the training data may be images of animals labeled as "dog", "cat", "bird", etc. The supervised learning model learns to predict the correct class for new inputs using the labeled examples in the training data.

In unsupervised learning, the training data is not labeled. The unsupervised learning model must therefore find structures and relationships in the data for itself. This can be used for tasks such as image segmentation, group detection and dimensionality reduction.

In the field of remote sensing, machine learning have been firstly used in several applications, such as change detection, LULC classification (or mapping), image preprocessing (e.g. fusion, segmentation etc.) or accuracy assessment. Support vector machine (SVM) and ensemble classifiers (e.g. Random forest) are mainly used due to the ability of the first method to handle high dimensionality data and ease of use for the second. Also, Random Forest achieved good results with a very low computational time [Belgiu et Drăguţ, 2016]. They are overtaking others methods used earlier in the field such as Linear Regression, Maximum Likelihood, K Nearest Neighbor and Regression Tree.

Random Forest is an algorithm of the ensemble classifier family, made of a set of tree-structured predictors. It uses a combination of multiple decision trees to make predictions. Each individual decision tree [Breiman, 1996] in a Random Forest is trained on a different subset of the data, and they all make their own predictions independently [Breiman, 2001]. The final prediction of the random forest is then determined by combining the predictions of all the individual decision trees. Random Forest is a powerful and effective machine learning method that can be used for both classification and regression tasks. It has been widely used for multiple remote sensing applications, such as LULC mapping [Pal, 2005, Colditz, 2015, Stefanski et al., 2013], and a lot of other tasks [Belgiu et Drăguţ, 2016].

Support Vector Machine is the second popular classifier for remote sensing applications. They are effective in dealing with high dimensional data and complex data sets. This makes them particularly useful for remote sensing applications, where large amounts

of data are collected by satellites and other sensors. It creates a hyperplane that maximally separates different classes of data. This allows for accurate classification and prediction by minimizing the misclassification rates [Mountrakis et al., 2011]. This method is very robust to overfitting problems, such as Random Forest, as a lot of regularisation parameters should be tuned by an expert [Cawley et Talbot, 2010]. This method has been very popular for LULC mapping [Kavzoglu et Colkesen, 2009].

2.2 Land Use Land Cover mapping: towards the use of deep learning

As classification problems become more and more complex (e.g. extraction of information that needs to be more and more accurate, taking into account large volumes of data etc.), deep learning methods have experienced a tremendous acceleration in the last few years due to their ability to process impressive amounts of data quickly and their capacity to learn complex features that have surpassed classical methods. As specified in section 2.1, deep learning models are composed of several layers capable of extracting several complex representations. Also, they can compute multiple representations from the representation in the previous layer which results in a hierarchy of data abstractions [LeCun et al., 2015].

2.2.1 Scene classification and semantic segmentation

In the field of LULC mapping, there are two mains types of deep learning applications, **scene classification** and **semantic segmentation**. Scene classification and semantic segmentation are two important tasks in the field of remote sensing image analysis. Scene classification involves assigning a label to an entire image, indicating what type of scene is depicted, such as "urban," "agricultural," or "forest." Semantic segmentation, on the other hand, involves assigning a label to each pixel in an image, indicating what object or feature is present at that location. Both scene classification and semantic segmentation can be important for a variety of applications, including land use and land cover mapping, environmental monitoring, and disaster management. For example, scene classification can be used to identify areas that are potentially suitable for renewable energy development or to track changes in land use over time. Semantic segmentation can be used to identify individual objects or features within an image, such as buildings, roads, or bodies of water, which can be useful for a variety of applications including urban planning and infrastructure management. Both of these applications are cover in the rest of the chapter.

In remote sensing, the word *classification* is often used to define automatic LULC map-

ping. On the other hand, in the deep learning community, *classification* (or scene classification in computer vision) consists in identifying the classes present in an image. A class assigned to a pixel is defined by the term *semantic segmentation* (Figure 2.2). For a better understanding, the concept of *semantic segmentation* will be used in the entire manuscript.

Figure 2.2 – *Difference between scene classification (a) and semantic segmentation (b). Reference data has been taken from Wenger et al. [2022a]. Legend is also in the paper.*

2.2.2 The emergence of Convolutional Neural Networks

Convolutionnal Neural Networks (CNNs) have received particular attention in recent years by the scientific community. They are well-suited for image recognition and processing as they perform mathematical operations called convolution to process the input data. CNNs are composed of multiple layers, including an input layer (e.g. images with 2 or more dimensions), one or more convolutional layers, a pooling layer, and an output layer. The input layer receives the raw input data, which in the case of image recognition is typically a matrix of pixel values. The convolutional layers apply filters to the input data to detect specific features, such as edges or shapes. The pooling layer then downsamples the output of the convolutional layers, reducing the dimensionality of the data and making the network more efficient. Finally, the output layer uses the processed data to make a prediction, such as the identity of an object in an image. During each convolution, feature maps are computed with a filter of a fixed-size kernel and the result is then passed through an activation function (e.g. ReLu, Sigmoïd, Leaky ReLu etc.). With this kind of approach, CNNs can detect local correlations.

The size of the kernel and the dimensions of the convolution filters are called hyper-

parameters. These are mostly not determined empirically and can be derived depending on the type of features to be extracted (e.g. spatial, spectral, spatial-spectral etc.) as well as the type of input data (e.g. time series, multispectral image, panchromatic image etc.). The main convolutional neural networks have either dimension 1 filters (1D-CNN), dimension 2 filters (2D-CNN) or dimension 3 filters (3D-CNN) (Figure 2.3). The only difference between these three networks is the size of the filter which varies from one to three dimensions.
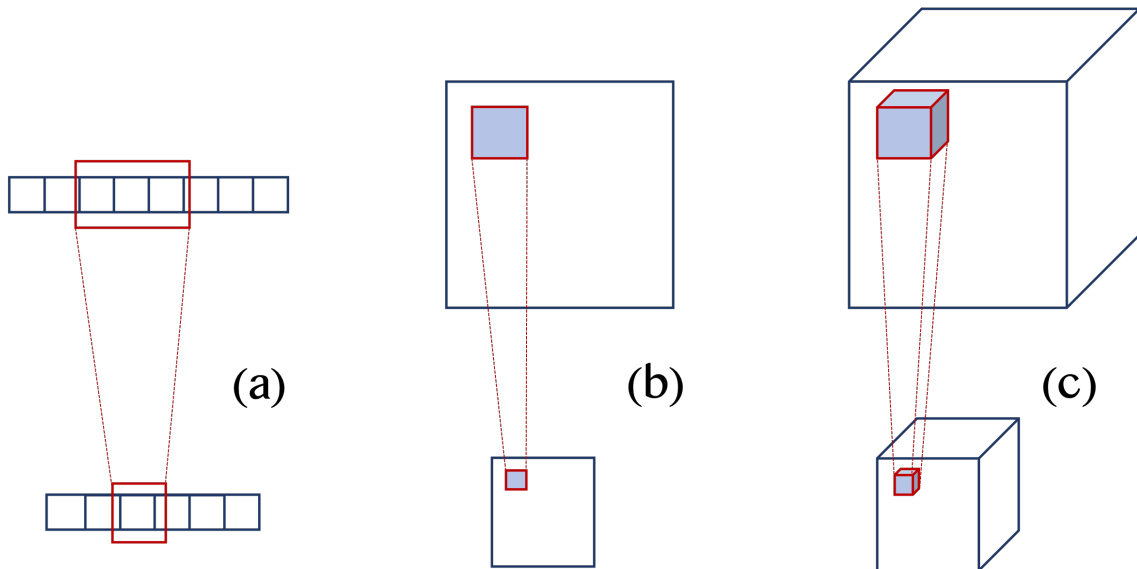
Figure 2.3 – *(a) One-dimensionnal CNN, (b) Two-dimensionnal CNN and (c) Three-dimensionnal CNN. Red lines represent the filter applied to the data.*

One dimensional CNNs have been mostly use for time series analysis but some works in the remote sensing community applied it to spectral [Pelletier et al., 2019, Song et al., 2019] and mostly hyperspectral imagery [Zhang et al., 2017, Mou et al., 2017]. The objective is to extract, from the multispectral series, a set of features allowing, from the spectral bands, the extraction of local patterns. Another work applied 1D-CNN to temporal hyperspectral imagery to distinguish season changes through spectral-temporal domain [Guidici et Clark, 2017].

Two-dimensional CNNs are the most commonly used types of CNNs for land cover classification from satellite images. 2D-CNNs allow the extraction of spatial features relating morphological attributes of the images to the desired reference classes. Since several years and the popularity of image understanding methods, many models (AlexNet [Krizhevsky et al., 2017], GoogleNet [Szegedy et al., 2015] or CaffeNet [Jia et al., 2014]) have been developed and tested by the remote sensing community for several applications, including LULC classification [Nogueira et al., 2017, Zhu et al., 2017a]. In order to consider the spatio-temporal aspect of the images, 2D-CNN are often paralleled with

1D-CNN, where the spatial and spectral features are extracted independently and then connected using a fully connected layer [Interdonato et al., 2019].

In order to take into account the multidimensional aspect of the images (both spectrally and temporally), the three-dimensional CNN has been developed by the scientific community. It can be used for the extraction of spatio-spectral [Li et al., 2017] but also spatio-temporal [Ji et al., 2018] features in order to offer LULC maps.

CNNs were an important breakthrough in the field of image segmentation [Iqbal et al., 2022] and achieved very promising results. However, a model brings even better results and has become, according to several works, a state of the art network in the field of segmented maps. U-Net is a convolutional neural network architecture that is commonly used for image segmentation tasks [Ronneberger et al., 2015]. The U-Net architecture (Figure 2.4) is based on encoder-decoder architecture, where an input image is first downsampled through a series of convolutional and pooling layers to extract features and create a compact representation of the image. This representation is then upsampled through a series of transposed convolutional layers, also known as deconvolutional layers, to produce a segmentation map that highlights the objects of interest in the input image. One key feature of the U-Net architecture is its use of skip connections, which concatenate the feature maps from the downsampling path with the upsampled feature maps from the upsampling path. This allows the network to retain spatial information from the downsampled representation, which can be useful for fine-grained segmentation tasks. U-Net has proven to be a powerful tool for medical image segmentation, particularly for tasks such as cell segmentation and tissue segmentation. In remote sensing, U-Net has been widely used for LULC mapping and has got very promising results [Li et al., 2018, Papadomanolaki et al., 2019a, Rakhlin et al., 2018].
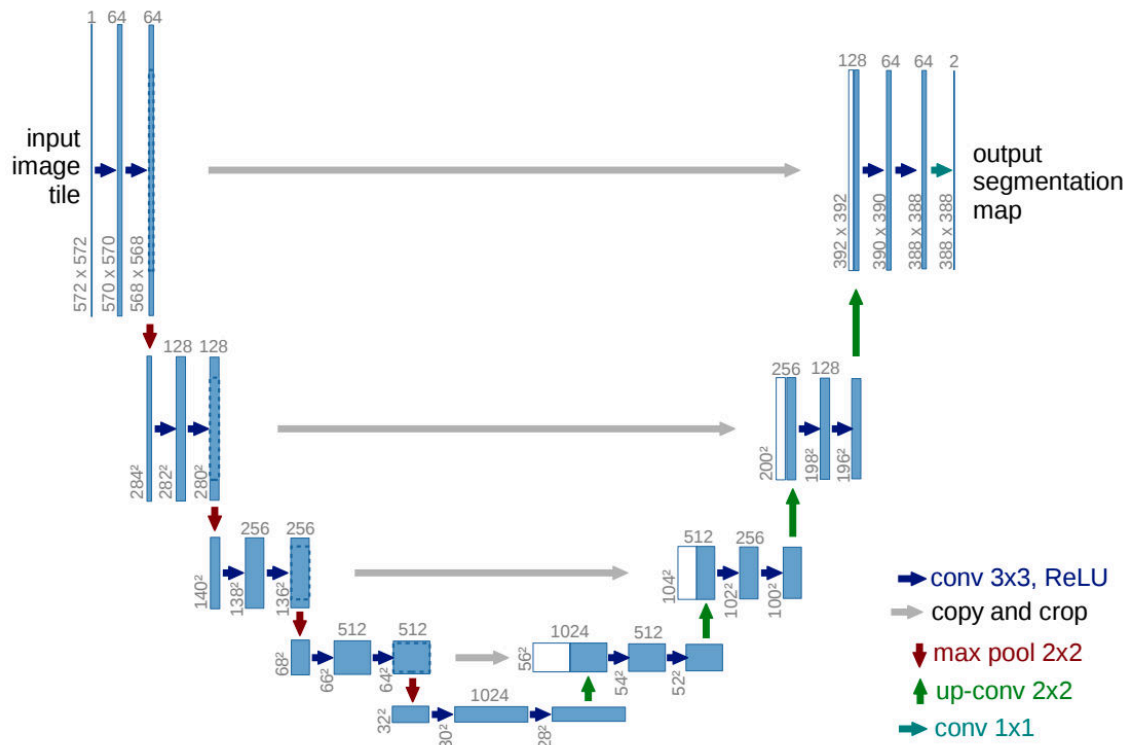
Figure 2.4 – *U-Net network architecture (Taken in Ronneberger et al. [2015]).*

2.3 MULTITEMPORAL AND MULTIMODAL DATA FOR SURFACE LAND USE LAND COVER MAPPING AND MONITORING

Remote sensing data come from multiple acquisition modes, SAR, multi-spectral sensors, aerial images, LiDAR and many more sensors. Also, satellites such as Sentinel have a short revisit which implies a large number of images. In this context, several techniques has been developed to deal with these various sources of data. This section covers the general scheme of multimodal or multisource data fusion, how they have been implemented in deep learning, and also which methods are employed to deal with multitemporal data (which are often four dimensionnal data in the case of remote sensing imagery).

2.3.1 Multimodal data integration

Remote sensing data fusion is not new and started to grow with the use of artificial intelligence methods where the need to combine multiple data sources was felt as they could improve LULC classification. In the past, many researchers investigates the fusion of heterogenous remote sensing data [Benediktsson et al., 2018, Chavez et al., 1991, Pohl et Van Genderen, 1998, Schmitt et Zhu, 2016]. The goal is to obtain a high-quality representation of the data [Zhang, 2010]. In order to perform data fusion in deep learning, images

should be paired with the same spatial footprint. The traditionnal methodology for data fusion is explained in figure 2.5 and contains three main stages : (a) Early stage or pixel level, (b) feature level or (c) decision level.
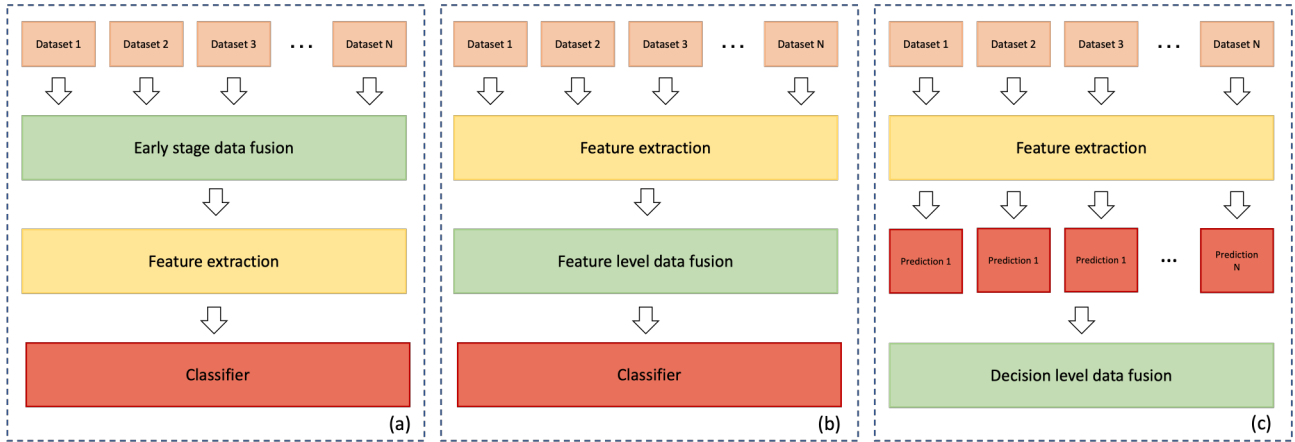
Figure 2.5 – *(a) Early stage or pixel level data fusion, (b) Feature level data fusion and (c) Decision level data fusion (Adapted from Ghassemian [2016])*

Early stage (or pixel level) fusion (Figure 2.5 (a)) aims to combine two or more images to obtain better quality information. The historical technique used is called *pansharpening* and consists in increasing the spatial resolution of the multi-spectral bands from the panchromatic image. In deep learning, this approach is called super-resolution and has outperformed the traditional technique for a few years [Huang et al., 2015, Yuan et al., 2018]. In deep learning, super-resolution also consists in improving the resolution of sensors that do not have a panchromatic band but between high-resolution sensors (e.g. Sentinel-2) and very high-resolution sensors (e.g. Spot-7) [Panagiotopoulou et al., 2021]. These methods are still popular and allow the artificial increase of the resolution of images from high temporal resolution sensors (e.g. Sentinel-2) by using very high spatial resolution imagery [Lanaras et al., 2018]. One of the major drawbacks of these methods is the loss of spectral information. Some works have used other networks (e.g. 3D-CNN) to increase the spectral quality of the data [Mei et al., 2017].

Feature level fusion (Figure 2.5 (b)) is one of the most used techniques for multi-modal remote sensing data fusion. The idea is to keep the data in its original state and then derive a set of features that are computed in parallel and put together in the classifier. From the classical machine learning methods, the feature level fusion methods were mainly executed by computing a segmentation image and then computing each feature derived from each data source [Dechesne et al., 2017, Fauvel et al., 2008]. In deep learning, this way of fusion data has been breakthrough in the recent years as multi-modal features can be extracted through several branches of a network [Xu et al., 2017, Chen et al., 2017]. In the field of

remote sensing and LULC mapping, feature fusion was perform using encoder-decoder like networks [Audebert et al., 2018], by merging branches of two streams CNNs with hyperspectral and multispectral imagery [Hu et al., 2017] but also with panchromatic and multispectral imagery [Gaetano et al., 2018].

Decision level fusion (Figure 2.5 (c)) consists of making predictions (or classifications) for each data source and making the final decision on the set of classifications or probabilities output from the models. In the case of deep learning, the duration of this type of fusion is significantly higher than the other two methods mentioned above because of the number of neurons in the output of the network. Audebert et al. [2018] and Piramanayagam et al. [2018] showed with their work that decision fusion is not the best approach to fuse multimodal data.

2.3.2 Dealing with Satellite Image Time Series

Recurrent neural networks (RNNs) are a type of artificial neural network that are particularly well-suited for processing sequential data, such as natural language, time series data, and audio. They are called "recurrent" because they use feedback connections that allow the network to process inputs that are dependent on previous inputs. RNNs have a "memory" element, called a hidden state, that allows them to retain information about past inputs. This allows the network to make predictions or decisions based on the entire sequence of inputs it has seen so far, rather than just the current input. There are several variations of RNNs, including long short-term memory (LSTM) and gated recurrent unit (GRU) networks. These variations use specialized layers to better capture long-term dependencies in the data, which is important for tasks such as language translation and language modeling.

In the field of remote sensing, this type of network has been widely used to deal with multitemporal satellite data (or sequential data). Some work uses 1D-CNNs and perform convolution over the temporal dimension. This is the case of TempCNN network [Pelletier et al., 2019] where authors made LULC classification in the South of France using Formosat-2 time series. This method does not cover the spatial aspect of the images and only extract spectro-temporal features. Pixel approches are suitable for natural areas classification and achieved promising results. In other works, LSTM and GRU have been combined with CNN to deal with the spatial and temporal aspect of the images : ConvLSTM and ConvGRU. Rußwurm et Körner [2018] used the ConvLSTM [Shi et al., 2015] and adapt it for remote sensing images. They demonstrated the high performance of this network for multitemporal satellite imagery. Ienco et al. [2019] used a network combining

convGRU (Convolutionnal Gated Recurrent Unit) and CNN to perform LULC classification of Reunion Island and Koumbia and compared this method with ConvLSTM and Random Forest classifier. Even if ConvGRU achieved higher results than other methods, ConvLSTM remains a very good method to obtain a classification from multi-temporal imagery. As mentioned in section 2.2.2, 3D-CNN can also handle multitemporal data as convolutions are made over spatial and temporal dimensions. However, ConvLSTM outperform 3D CNN for remote sensing applications [Chang et al., 2022].

2.4 Conclusion

In this chapter, we presented most of the existing methods for LULC mapping based on multitemporal and multimodal deep learning approaches. Semantic segmentation and pixel-based approaches are the most common ones used in this domain. ESA World Cover and OSO use a classical machine learning pixel approach (Random Forest). This one, adapted for natural areas, leads to strong salt and pepper effects when applied to urban areas. For this, deep learning pixel approaches (e.g. TempCNN) discriminate the UF into a single class. Thus, for all the works of this thesis, it was chosen to use semantic segmentation approaches, more adapted to urban areas and demonstrated by some existing works [El Mendili et al., 2020]. The strength of deep learning also resides in its ability to merge features within a single network, as opposed to traditional machine learning approaches, particularly due to the more complex scaling. In order to assess if deep learning approaches are relevant, it is necessary to have adapted training datasets.

MULTISENGE : A MULTIMODAL AND MULTITEMPORAL BENCHMARK DATASET FOR LAND USE/LAND COVER REMOTE SENSING APPLICATIONS

3

CONTENTS

Today, multitemporal and multimodal Land Use Land Cover datasets are still not easily available, especially for semantic segmentation applications. The contribution of multimodal data in the classification of LULC has already been demonstrated. One of the challenge in the remote sensing community is to find a way to combine optical and radar

imagery to perform LULC classification. These sensors bring on the one hand knowledge on materials through optical imagery and on the other hand structural characteristics through radar imagery.

In order to propose a valuable dataset in term of coverage and in term of accuracy in semantic classes, an existing topographic LULC database covering the Grand-Est region has been used to build a benchmark dataset (MultiSenGE). This one takes as input optical and SAR time series (Sentinel-1&2). To address the issue of spatial resolution of the sensors used, a set of methods has been developed to obtain thematic consistency between satellite and LULC data. The first tests were performed on optical mono-temporal data to validate the dataset.

> This chapter has been published in **ISPRS Annals** [Wenger et al., 2022a] and presented at **ISPRS 2022** in Nice (France) as well as the publication of the dataset on Zenodo [Wenger et al., 2022b].

3.1 INTRODUCTION

The constant evolution of earth observation missions has allowed the acquisition of a large amount of satellite data. They have considerably changed the way humanity manages its territory. One of the major examples is the Copernicus program developed by the European Space Agency (ESA) which consists in the deployment of several constellations of satellites to monitor the Earth Surface. This is the case with the Sentinel missions, which currently consists of six missions designed as a two-satellite constellation. Sentinel-1 and Sentinel-2 provide open-access and freely Synthetic Aperture Radar (SAR) and multispectral imagery with a very short revisit time, respectively 5 days for Sentinel-2A/2B and between 6 and 12 days depending on the location on Earth for Sentinel-1A/1B.

The high revisit period of these two sensors allows the use of time series to study the dynamics and evolution of processes and objects of interest. Satellite Image Time Series (SITS) have already been used for the analysis of agricultural areas [Bégué et al., 2018, Bellón et al., 2017, Kussul et al., 2017, Rußwurm et al., 2020], forest areas [Wulder et al., 2012, Pickell et al., 2016] or classification of Land Use/Land Cover (LULC) [Inglada et al., 2017]. In addition, the joint use of SAR and optical data continues to increase the interest of researchers, especially in the case of LULC mapping [Ienco et al., 2019, Steinhausen et al., 2018]. This combination has already shown its performance in many other works such as the detection of natural areas [Dusseux et al., 2014, Mngadi et al., 2021], the detection of changes [Gao et al., 2017] or the mapping of urban areas [Iannelli et Gamba, 2018].

To deal with this increasing amount of data, new techniques based on neural networks have been developed and offer promising results in the classification of LULC from multiple data sources [Ma et al., 2019]. Multimodal and multitemporal datasets are currently quite rare and remain for the most part specific to applications (Object detection, Scene classification, Semantic segmentation or Instance segmentation) and it is important to have a variety of data sets for the application of deep learning models. To our knowledge, only two datasets use Sentinel-1/Sentinel-2 pairs for scene classification or semantic segmentation, BigEarthNet [Sumbul et al., 2021] and SEN12MS [Schmitt et al., 2019b]. BigEarthNet proposes 590,326 pairs of single-time annotated images with the Corine Land Cover reference data over 10 countries of Europe. SEN12MS offers 180,662 Sentinel-1 and 2 triplets and MODIS Land Cover over several regions in the world and for spring, summer and winter seasons to perform semantic segmentation.

In order to support the lack of multimodal and multitemporal datasets, we decided to produce MultiSenGE, a benchmark dataset covering the Grand-Est region in France (57,433 km^2 which represents 10.6% of the French territory). The objective is to focus the bench-

mark dataset on semantic segmentation and classification. This dataset offers Sentinel-1, Sentinel-2 and LULC triplets, a land cover data recently available with an open-source licence over this territory. Compared to other existing datasets, MultiSenGE allows to classify urban surfaces into 5 LULC classes, against only 1 for SEN12MS using MODIS Land Cover as reference data and 11 for BigEarthNet using CORINE Land Cover as reference data. We use OCSGE2-GEOGRANDEST which have the advantage to own a Minimum Mapping Unit (MMU) less than $50m^2$, which gives it a higher geometric acuracy than existing LULC products and is close and consistent with the spatial resolution of Sentinel imagery (10m).

First is presented the satellite and LULC data on the study area used to build the dataset. Then the methodology to process the reference data and to create triplets of patches is presented. Finally, baseline results performed on the *beta* version of MultiSenGE only based on urban thematic classes are described before conclusion and perspectives.

3.2 REFERENCE DATA AND SATELLITE IMAGERY

3.2.1 Study sites and reference data

The dataset covers a large territory (57,433 km^2) in eastern France and corresponds to an administrative French district (Figure 3.1) extended from Alsace in the East to the Ardennes and Marne in the West. This area have been chosen due to the availibility of a new, accurate and up-to-date vector LULC database named OCSGE2-GEOGRANDEST (www.geograndest.fr). This open-access vector database[1] was built by visual interpretation of aerial photographs for 2019/2020. It is organized into four levels of nomenclature where the first level categorizes land cover into four classes (1) artificial surfaces, (2) agricultural areas, (3) forest areas, and (4) water surfaces. At the most accurate level (1:10,000), 53 LULC classes map the region and the size of the smallest elements is $50m^2$. In order to obtain a generic reference data with 14 classes and have class consistency at 10m spatial resolution, several preprocessing steps are performed on the original topographic vector database (see 3.1). In the OCSGE2 layer, all the roads have the same degree of importance. Some of them are too small to be distinguished at 10 m spatial resolution. Then, in order to produce an adapted road network label, a second database (BDTOPO-IGN), produced by IGN describing lines in vector format, with the degree of importance, is pre-processed.

---

[1] (Etalab v2 Licence)
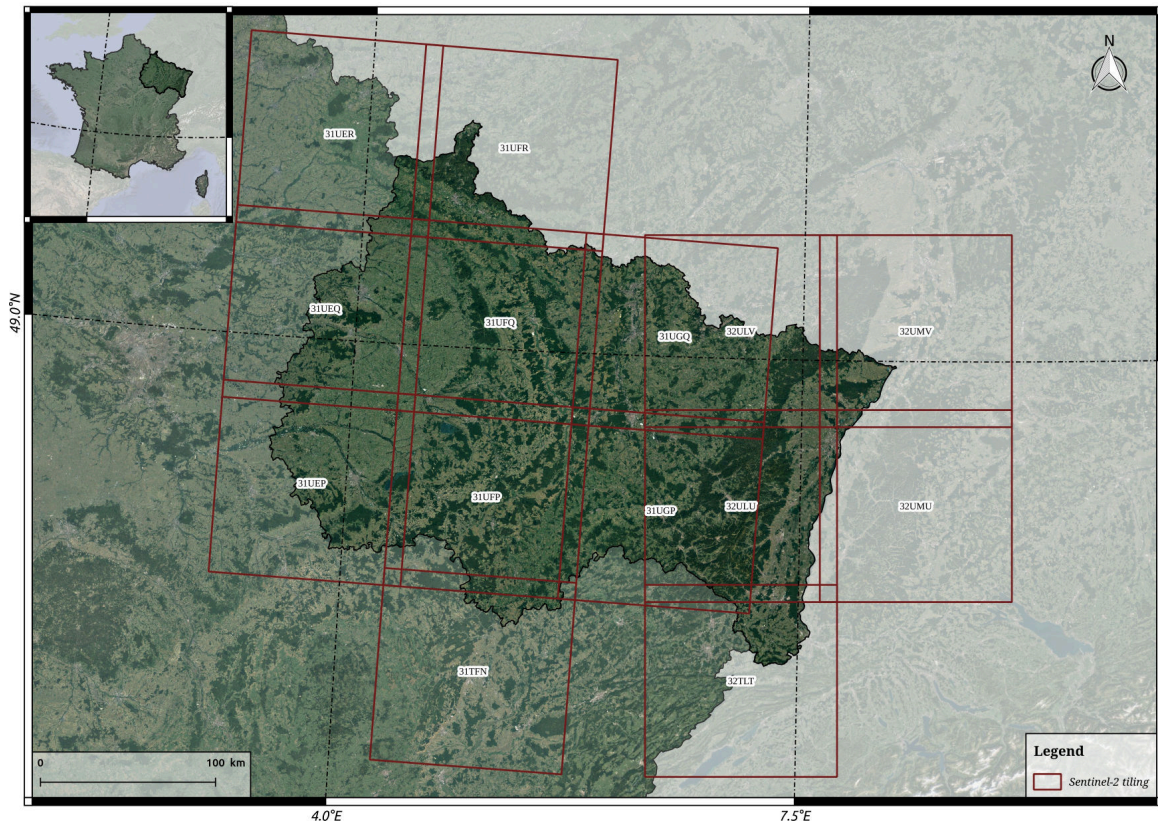
Figure 3.1 – *Study area with Sentinel-2 tiling grid.*

3.2.2 Sentinel-2

Sentinel-2 offers 13 spectral bands at 10m and 20m spatial resolution. The Satellite Image Time Series (SITS) for 2020 (14 tiles - Figure 3.1) are downloaded from the Theia land services and datacenter (https://www.theia-land.fr/), in L2A format, corrected from atmospheric effects with a cloud mask (surface reflectance product). Only free cloud cover images (with cloud cover strictly less than 10%) are automatically downloaded by querying their database using the script *theia_download* [Hagolle, 2021]. For the MultiSenGE dataset the 10 spectral bands at 10 meters (B2, B3, B4, B8) and 20 meters (B5, B6, B7, B8A, B11 and B12) spatial resolution are used.

3.2.3 Sentinel-1

Sentinel-1 is equipped with C-band SAR sensors wich allows the acquisition of imagery day and night without weather disturbances compared to optical imagery. It provides data in dual polarization with two product types : Grand Range Detected (GRD) and Single Look Complex (SLC) Filipponi [2019]. GRD products, used to construct this dataset, consist on SAR data that have been multi-looked and projected to ground range using Earth ellipsoid model.

The SAR images available in ascending and descending orbits for 2020 were downloaded and pre-processed using the S1-Tiling [Koleck et Centre National des Etudes Spatiales (CNES), 2021] processing chain developed by CNES (Centre National d'Etudes Spatiales). This processing chain can be divided into four points : (1) automatic downloading of Sentinel-1 data thanks to the EODAG library which offers the possibility to request different servers to always have data available on the studied area, (2) slicing of the SAR data according to the Sentinel-2 tiling, (3) orthorectification of the newly sliced SAR scenes and (4) application of a multi-temporal filter to reduce the speckle and preserve the spatial information.

3.3 MultiSenGE production

3.3.1 Reference data processing

In order to obtain LULC data at 10 meters spatial resolution, five pre-processing steps are applied on the reference data (Figure 3.2). The five steps consists in (1) resampling reference labeled data, (2) removing the smallest polygons thanks to the connected component labeling method applied on each selected class, (3) filling the holes resulting from this method by nearest neighbor, (4) applying a mathematical morphology of closure to smooth the outlines of each class on the final data and (5) adding roads from the second database.
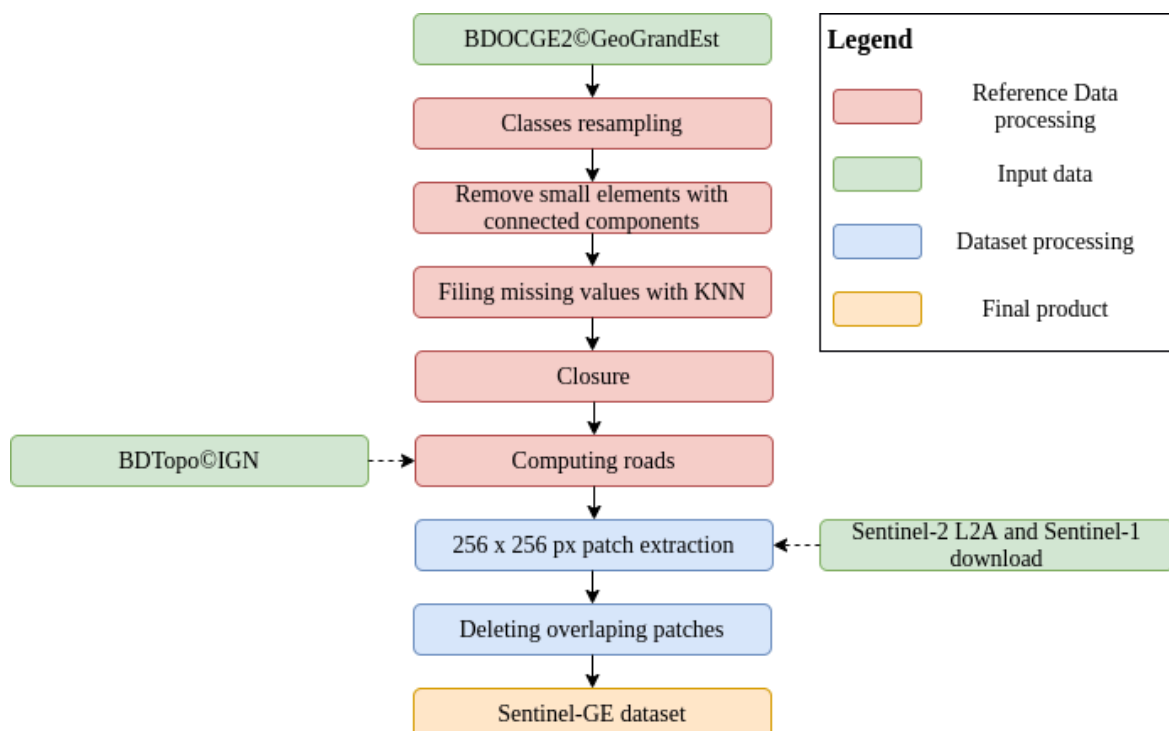
Figure 3.2 – *Methodology for MultiSenGE production.*

First, to obtain a number of labeled classes adapted to the spatial resolution, we kept the classes at level 3 of the nomenclature with some semantic reclassification to reduce the complexity of the typology, especially for semantic segmentation applications. Finally, the final LULC typology includes 14 classes with 5 classes for urban areas and 9 classes for natural surfaces (Table 3.1).

Original level 1 typology	New level 3 typology
Urban Areas (1)	Dense Built-Up (1)
	Sparse Built-Up (2)
	Specialized Built-Up Areas (3)
	Specialized but Vegetative Areas (4)
	Large Scale Networks (5)
Agricultural areas (2)	Arable Lands (6)
	Vineyards (7)
	Orchards (8)
	Grasslands (9)
	Groces, Hedges (10)
Forests and semi-natural areas (3)	Forests (11)
	Open Spaces, Mineral (12)
Wetlands (4)	Wetlands (13)
Water Surfaces (5)	Water Surfaces (14)

Table 3.1 – *Class name and pixel value in the final reference data in a raster format.*

The second step is applied to reduce the spatial complexity of the reference data which can contain small polygons ($50m^2$) that will add noise to the database and are not visible at 10 m. This important step is based on a connected components method to extract the smallest polygons and then sort them according to their area. All polygons smaller than 2.5 ha (total surface of 250 pixels) for each class are left blank, which is more accurate than the MMU of some existing LULC products such as Corine Land Cover[2]. This value is an empirical choice after several tests. In the third step, a nearest neighbor method is applied to fill the holes. It consists in finding, using Euclidean distance and the nearest neighbors, the value of the missing data by calculating the mean of their value [Troyanskaya et al., 2001]. A higher weight is assigned to the nearest neighbors of the missing data. A closure with a rectangle morphological object is then applied in the fourth step to smooth the

---

[2]https://land.copernicus.eu/pan-european/corine-land-cover

new level 3 reference data (Figure 3.3). Finally, large Scale Networks are added in post-processing and come from OCSGE2-GEOGRANDEST for the railway and BDTOPO-IGN for the most important road networks. A buffer of 30 meters is applied for highways and 10 meters for the second most important roads or railways, which is consistent with the 10 m spatial resolution. After merging all the vector data in a unique vetor layer, all polygons are rasterized at 10 m spatial resolution.
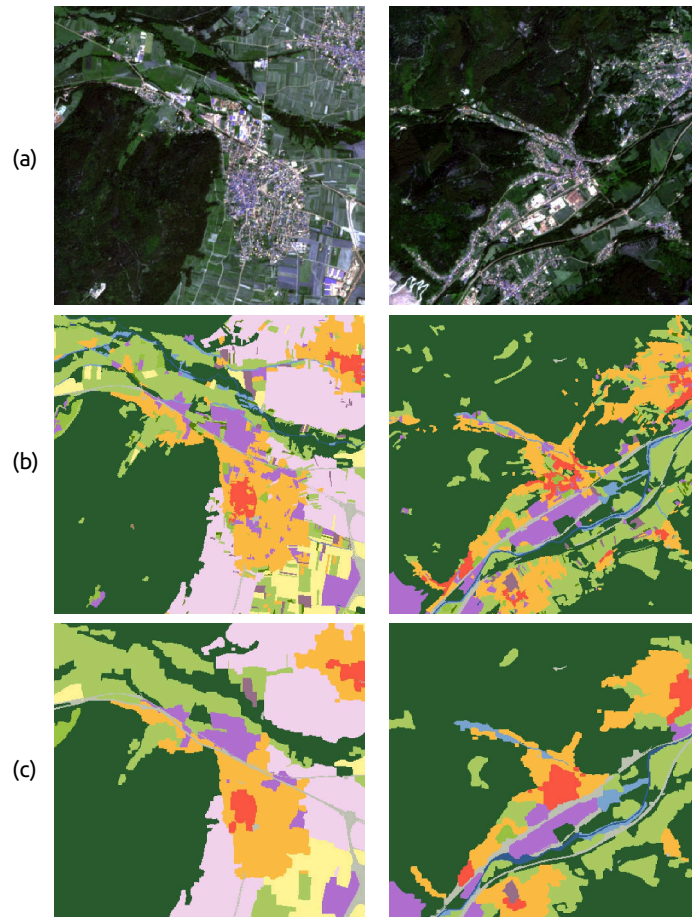
Figure 3.3 – *Comparison between (a) Sentinel-2 image in RGB, (b) reference database before the application of the connected components method and the (c) final reference data (Color legend like in Figure 3.4)*

3.3.2 Triplet data preparation

The Sentinel-1 SAR SITS, Sentinel-2 optical SITS and the reference data are cut into 256 x 256 pixel patches. The VV and VH bands from the pre-processed Sentinel-1 images are stacked for each patch. For Sentinel-2, the bands at 10 meters are kept and the bands at 20 meters are resampled to 10 meters spatial resolution using cubic interpolation to have homogeneity in the final data. The 10 bands are then stacked for each patch. Finally, the simplified LULC reference data is also cut following the same footprint to build triplets with dual-pol Sentinel-1 image patches and multispectral Sentinel-2 image patches. This

results in triplets containing the reference data, the Sentinel-2 time series (n number of dates for each patch) and the Sentinel-1 time series (m number of dates for each patch). In a post-processing step, overlapping patches are removed to obtain a spatial independence of all patches contained in the dataset. This overlap region mainly concerns the 10km supperposition area between adjacent tiles.

3.3.3 Structure of the benchmark dataset

Each triplet is described by labels to perform both scene classification and semantic segmentation (Figure 3.5). In addition to the classes, the *GeoJSON* file also contains the names of all associated Sentinel-2 and Sentinel-1 patches as well as the specific projection for each patch. The projection of each patch is in UTM/WGS84 format inherited from its original tile. For the current version, the dataset contains 8,157 non-overlapping triplets along with the *GeoJSON* file containing all the classes present in the patch. The mosaic level 3 reference data product and its typology (Figure 3.4) will also be available for download.

Dense Built-Up
Sparse Built-Up
Specialized Built-Up Areas
Specialized but Vegetative Areas
Large Scale Networks
Arable Land
Vineyards
Orchards
Grasslands
Groves, Hedges
Forests
Open Spaces, Mineral
Wetlands
Water Surfaces
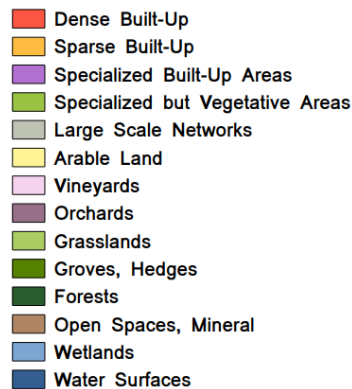
Figure 3.4 – *Typology and colors proposed for MultiSenGE.*

MultiSenGE contains four folders with (1) the simplified reference data patches, (2) the Sentinel-2 patches, (3) the Sentinel-1 patches and (4) the *GeoJSON* files. The files contained in the folders are identified according to the following nomenclatures:

- *ground_reference* : {tile}_GR_{x-pixel-coordinate}_{y-pixel-coordinate}.tif

- *s2* : {tile}_{date}_S2_{x-pixel-coordinate}_{y-pixel-coordinate}.tif

- *s1* : {tile}_{date}_S1_{x-pixel-coordinate}_{y-pixel-coordinate}.tif

- *labels* : {tile}_{x-pixel-coordinate}_{y-pixel-coordinate}.json

where *tile* is the Sentinel-2 tile number, *x-pixel-coordinate* and *y-pixel-coordinate* are the coordinates of the patch in the tile and *date* is the date of acquisition of the patch image as

the time series is extracted from each sensor. *GR* means Ground Reference and correspond to the ground reference patches, *S2* Sentinel-2 and correspond to the Sentinel-2 patches and *S1* Sentinel-1 for the Sentinel-1 patches. Users will find 72,033 multi-temporal patches for Sentinel-2 and 1,012,227 multi-temporal patches for Sentinel-1. To facilitate the reading and extraction of dates from the dataset, tools have been developed and are available on a code hosting service.
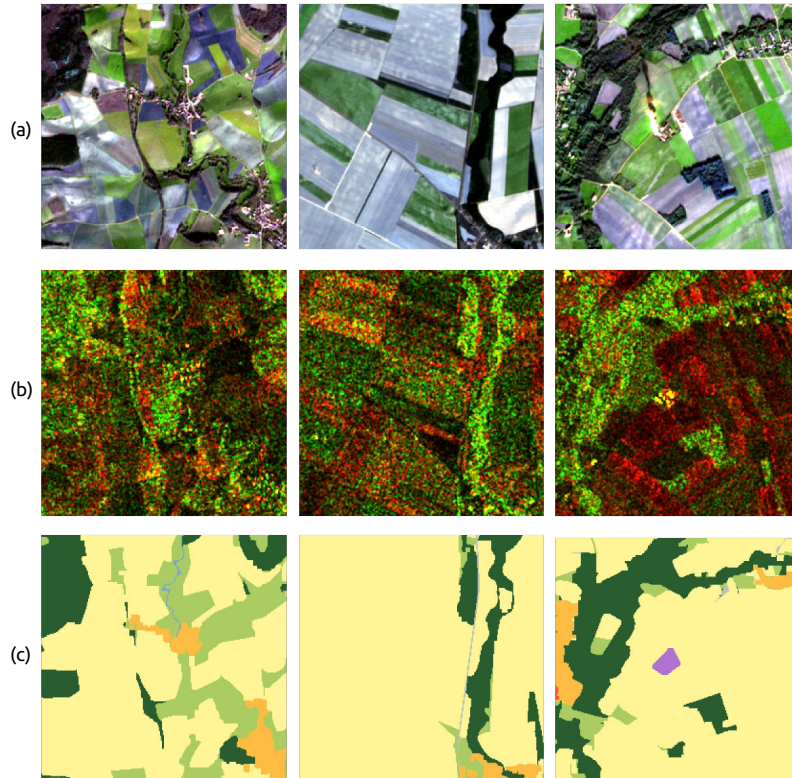
Figure 3.5 – *Triplets patches example with mono-temporal (a) Sentinel-2 in RGB, (b) SAR Sentinel-1 (Red : VV, Green : VH) and (c) the ground reference associated.*

3.4 Baseline results

First baseline results are performed on urban areas using only single-time patches and the five urban classes of the dataset (Table 3.1, classes 1 to 5). First experiments focused on the Moselle department (T31UGQ). We then selected a subset on the tile for training and validation steps and a different spatially split subset for testing zone [Saraiva et al., 2020]. Two multiclass U-Net networks with VGG-16 as backbone are tested (Table 3.2). U-Net-IRRG uses the 3 IRRG bands (InfraRed, Red and Green) as well as weights pre-trained on ImageNet [Deng et al., 2009] and U-Net-Index uses the 3 IRRG bands as well as 3 spectral and textural indices, the NDVI (Normalized Difference Vegetation Index), the NDBI (Normalized Difference Building Index) and the entropy [Haralick et al., 1973] computed

on the NDVI (*eNDVI*). The weights of U-Net-Index have been randomly initialized. As an imbalanced dataset, a weighted categorical cross-entropy loss was used by assigning higher weights to the less represented classes [Audebert et al., 2018]. This represents the inverse of the class frequency. The city of Metz (Grand Est, France) has been selected as a test area for the urban areas semantic segmentation application. Class 6 represents the aggregation of the other non-urban classes (Classes 6 to 14 in Table 3.1).

Parameters	Networks	
	U-Net-IRRG	U-Net-Index
Backbone	VGG-16	VGG-16
Pretrained	ImageNet	*Random*
Batch size	8	8
Loss	wCCE	wCCE
Optimizer	Adam	Adam
Initial LR	0.0001	0.0001
Bands	IRRG	IRRG + Index

Table 3.2 – *Networks selected for single-time application on urban areas.*

We selected 80% of patches for training and 20% for validation, outside the training area. Each network was trained for 100 epochs with Adam as the optimizer for gradient descent. Adam was prefered to SGD (Stochastic Gradient Descent) because it is more stable than the latest for semantic segmentation. Two metrics, classically used for multi-class classifications, were selected, a $F1_{Score}$ weighted according to the frequency of each class within the test area in order to have an overall metric for each selected network (Table 3.3) and an unweighted $F1_{Score}$ to have a statistical evaluation of each class (Table 3.4).
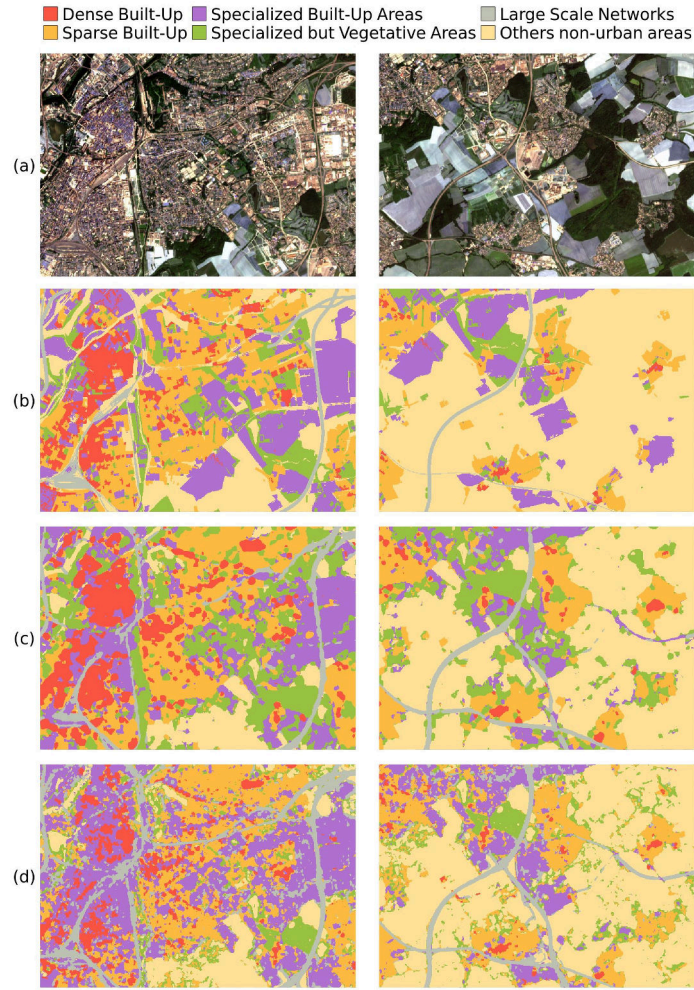
Figure 3.6 – *Visual results for two subsets over Metz. (a) is the Sentinel-2 RGB, (b) is the ground reference in a raster format, (c) U-Net-IRRG prediction and (d) U-Net-index prediction.*

Weighted $F1_{Score}$	
U-Net-IRRG	U-Net-Index
0.7364	0.7214

Table 3.3 – *Weighted $F1_{Score}$ for baseline results over Metz.*

The statistical results show better scores for the U-Net-IRRG method compared to the U-Net-Index method with weighted $F1_{Score}$ of 0.7364 and 0.7214 respectively (Table 3.3). Moreover, four of the six classes studied have a better $F1_{Score}$ for U-Net-IRRG than U-Net-Index (Table 3.4). These statistical results are confirmed by the visual results presented in Figure 3.6. We notice a better homogeneity in the classification of all the classes on the test area for the first method while the second method offers a much more fragmented classification with an underestimation of several classes.

These baseline models are the first ones applied only on a single tile (T31UGQ).

Classes	$F1_{Score}$	
	U-Net-IRRG	U-Net-Index
Class 1	**0.5199**	0.4822
Class 2	**0.6565**	0.5864
Class 3	0.5090	**0.5372**
Class 4	**0.3141**	0.2048
Class 5	**0.5862**	0.4295
Class 6	0.8549	**0.8579**

Table 3.4 – *Classes F1$_{Score}$ for baseline results over Metz.*

3.5 CONCLUSION

This paper presents MultiSenGE, a new large scale multimodal and multitemporal benchmark dataset covering a large area in the eastern of France. It contains 8,157 triplets of 256 x 256 pixels of Sentinel-1 dual-polarimetric SAR data, Sentinel-2 multispectral level 2A images and a LULC reference database built to obtain a typopoly adapted to the 10m spatial resolution for year 2020. Moreover, the proposed 14 classes are consistent to map French or European national-scale landscape. Triplets data offer users the possibility to perform semantic segmentation and scene classification on multitemporal and multimodal data with large input patches. The baseline results have shown its ability to offer encouraging first results on urban areas classification with patches only with the single-time optical imagery. Others tests are ongoing and should improve these first results by using multitemporal and multimodal imagery offered by this dataset as well as different deep learning techniques. This benchmark dataset for LULC remote sensing applications will soon be shared with the scientific community through the Theia Data and Services center[3]. It also could be enriched by additionnal data as complete Sentinel-2 SITS with cloud and snow masks.

---

[3]https://www.theia-land.fr/pole-theia-2/

U-Net feature fusion for multi-class semantic segmentation of urban fabrics from Sentinel-2 imagery: an application on Grand Est Region, France

4

Contents

Deep learning approaches have been used for several years for the classification of several LULCs and more specifically UF. In the previous chapter, the interest of proposing a multimodal and multitemporal dataset from Sentinel-2 optical and Sentinel-1 radar imagery has been demonstrated. The first mono-temporal experiments on the UF in 5 classes have shown first interesting results. In order to continue in this direction, fusion methods between Sentinel-2 single-date optical imagery and spectral and textural indices have been explored with the objective of improving UF mapping. In this chapter, a beta version of MultiSenGE was used which only covers five districts in the Grand-Est region over two Sentinel-2 tiles (*Bas-Rhin*, *Moselle*, *Vosges*, *Meurthe-et-Moselle* and *Haut-Rhin*). At the moment of this research, only five departements of the OCCGE2 database was already produced. The entire region was only available at the begining of 2022. In this context, this paper answers the first objective which was to evaluate the performance of deep learning approaches for Urban Fabric mapping.

This chapter has been published in **International Journal of Remote Sensing** [Wenger et al., 2022c].

4.1 INTRODUCTION

By 2050, more than three out of four people will live in cities. For comparison, in 1950, it was only 33% of the population that lived in cities [United Nations Department of Economic and Social Affairs Population Division, 2018a], slightly more than one in four inhabitants. As a consequence, urban areas are increasing as a result of development of built-up areas, network infrastructure, industrial areas or other built-up areas. This urban sprawl triggers changes in landcover with the consumption of agricultural and natural areas, and has impacts on the ecosystems with important ecological, climate and social transformations [Irwin et Bockstael, 2007, Zhu et al., 2019]. Most studies quantify the dynamic of urban footprint [Puissant et al., 2011] which includes the road network, buildings, vegetation, and impervious surfaces [El Mendili et al., 2020]. Few of them analyses the inner dynamics of urban areas through the changes of urban fabrics (UF) corresponding to a specific spatial organization of basic components of the city. Several works based on geographic object-based image analysis (GEOBIA) have been explored to obtain land cover land use (LULC) classifications [Souza-Filho et al., 2018, De Luca et al., 2019, Uddin et al., 2018] with a higher accuracy than "per-pixels" methods, which are considered insufficient and do not take into account the neighbors of each pixel.

With the multiplication of Earth Observation satellites, the amount of acquired satellite images continue to grow exponentially. The evolution of computing power in computer science has motivated researchers to develop classification methods based on neural networks [Kamga et al., 2021] and running on GPUs instead of CPUs. Ma et al. [2019] demonstrate the renewed interest in these techniques by conducting a review using these neural network-based deep learning methods in remote sensing. Deep learning covers several fields in remote sensing, whether for image fusion (example of pan-sharpening) [Xing et al., 2018], scene classification and object detection [Zhong et al., 2018, Ding et al., 2018, Sumbul et al., 2019], LULC classification [Marcos et al., 2018, Zhu et al., 2018] or semantic segmentation [Chen et al., 2018, Kemker et al., 2018]. These fields can be grouped into four main tasks : image preprocessing, change detection, accuracy assessment and classification [Ma et al., 2019].

Semantic segmentation methods are used in several domain applications from medical image segmentation to object detection on photographs [Han et al., 2019, Shin et al., 2016]. In computer vision, CNN (Convolutional Neural Networks) are excellent networks to analyze images containing high level spatial features. CNNs, along with VGG networks, are used in remote sensing for slums detection [Wurm et al., 2019], object detection such as aircraft [Ding et al., 2018] or change detection [Amin Larabi et al., 2019]. Moreover,

encoder/decoder networks provide a higher accuracy than classical CNNs in detecting boundaries between objects [Chhor et Aramburu, 2017]. U-Net [Ronneberger et al., 2015] and SegNet [Badrinarayanan et al., 2017], two encoder/decoder networks have been developed for image classification. U-Net has shown excellent results in biomedical image segmentation and SegNet in scene classification. These networks have also been used in remote sensing for land cover classification of very high resolution images [Zhang et al., 2018] but also in the classification of UF using high resolution Sentinel-2 RGB imagery and pretrained ImageNet [Russakovsky et al., 2015] weights [El Mendili et al., 2020]. Sefrin et al. [2021] used an encoder/decoder like network combined with a LSTM (Long Short-Term Memory) to classify land cover into 8 classes from Sentinel-2 images and obtained high classification score in several land use classes.

Transfer learning [Lu et al., 2015] is also used in many semantic segmentation works. This technique consists in assigning the weights of a pre-trained network on a data source to the target network that we intend to train [Oquab et al., 2014]. This result is a time saving in the training of the target network and also allows to counterbalance a dataset containing very few entries [Xie et al., 2016, Momeni et al., 2016]. Kemker et al. [2018] showed the efficiency of this method by transferring weights from a source network trained on panchromatic images to a network treating multispectral data. Shendryk et al. [2019] trained a network for scene classification using planetescope imagery and then transferred the weights for Sentinel-2 image classification. Iglovikov et Shvets [2018] modified the decoder part of a U-Net in VGG11 to be able to use pre-trained weights on ImageNet [Russakovsky et al., 2015]. Thanks to this fine-tuning technique, they obtained high-quality results that can be further improved by using deeper networks such as VGG16 or any other ResNet-like networks.

Many studies have shown the interest of fusion techniques between two networks. For instance, Audebert et al. [2018] trained a modified SegNet using feature fusion method to detect UF classes from aerial images. They fuse the encoder phase from two SegNet networks, one using exogenous indexes and the other IRRG (Infra-Red, Red and Green) bands using pretrained ImageNet weights. They obtained better results using this new method. Network fusion is also applied when using CNN and ConvGRU (Convolutional Gated Recurrent Unit). Ienco et al. [2019] combined these two networks and two data sources for land cover mapping of Reunion Island and Koumbia and shows the efficiency of networks fusion from different sources. Hu et al. [2017] developed a two stream CNN for LULC classification using radar and hyperspectral images. In addition, many datasets have been developed to perform research on multi-modal fusion, whether between opti-

cal and radar imagery [Schmitt et al., 2019a, Sumbul et al., 2021], aerial imagery and a Digital Surface Model (Vaihingen and Potsdam datasets developed for ISPRS 2D Semantic Labeling Challenge[1]) or hyperspectral and LiDAR imagery (Houston2013 dataset [2]).

The use of spectral and textural indexes when applying neural networks are increasingly used as an external addition of exogenous indexes calculated with the spectral bands of the sensor chosen for the study. These hand-crafted features allow for differentiation of different types of spatially structuring objects to accelerate network learning and improve classification results [Liu et al., 2017]. Campos-Taberner et al. [2020] developed a method to determine the importance of Sentinel-2 bands and spectral and textural indices in a neural network. They noted that NDVI (Normalized Difference Vegetation Index), NIR (Near Infra-Red) and red bands, and entropy calculated on NDVI (*eNDVI*) are the features providing the most relevant information to the network. NDVI index is frequently used in the detection of UF because it allows to accentuate the distinction between these spaces and the vegetative areas due to the important difference in the spectrum of the materials constituting them. In fact, it is widely used in land cover/land use detection research [Ienco et al., 2019, Inglada et al., 2017]. NDBI (Normalized Difference Building Index) [Zha et al., 2003], is an index developped to rapidly extract urban fabric [Yi et Jianhui, 2016]. It works like the NDVI by making a combination of different spectral bands, in this case for this index the MIR (Mid Infra-Red) and the NIR (Near Infra-Red).

In this context, our research focused on the contribution of feature fusion from Sentinel-2 high spatial resolution imagery to map UF based on a generic typology for France. In addition, this study aims to show that the use of exogenous indices (NDVI, NDBI and *eNDVI*) can improve the classification results of these UF. The paper is structured in five sections. Section 2 will describe the materials and methods. Section 3 will present the results for each method and study area. In section 4, we will discuss the results presented previously before concluding and developing our research perspectives in section 5.

4.2 MATERIALS AND METHODS

In this section, the study sites is firstly described (section 2.1) followed by the presentation of both satellite and databases processed to obtain a reference dataset allowing to improve the UF map into five thematic classes (section 2.2). The proposed workflow is then explained to produce urban semantic segmentation based on Sentinel-2 mono-date image.

---

[1] http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html
[2] https://hyperspectral.ee.uh.edu/?page_id=459

4.2.1 Study Sites and Test areas

The 'Grand Est' region is an administrative French zone extended from Alsace in the East to the Ardennes and Marne in the West, covers 57,441 km^2 as well as many large urban areas such as Strasbourg, Metz, Nancy and Reims. Among the sixteen tiles covered by Sentinel-2, two of them are chosen for study sites for training and testing our classification models from imbalanced reference data (Figure 4.1). The land cover classes distribution of both tiles is representative of the whole region covered urban, peri-urban and rural areas with a diversity of UF. Three subsets of test with a gradient of urbanisation are selected in order to assess the robustness and precision of the methods tested for a diversity of case studies. The first test area is located on tile 32ULU North, one the most important city of the East French Region (Strasbourg with more than 500 000 inhabitants), the second one includes part of the city of Metz (one of the four cities with around 200,000 inhabitants) and the last one is located near the smaller city of Saint-Avold (tile 31UGQ) and is representative of cities with less than 50,000 inhabitants (Figure 4.1).
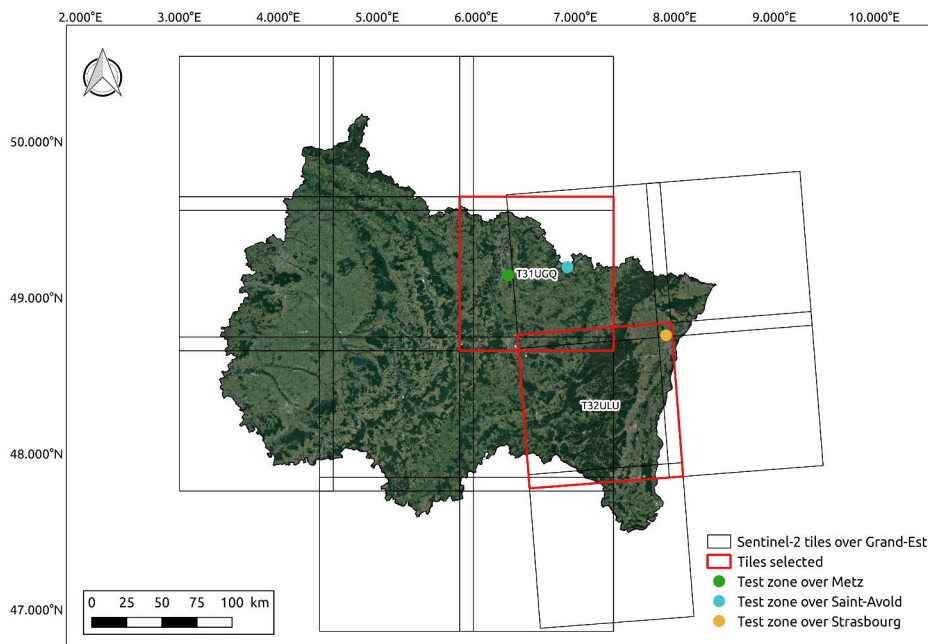
Figure 4.1 – *Grand Est region, France, including Sentinel-2 tiles and test areas where the experiments were made. (Coordinate system used is World Geodetic System 1984/EPSG4326)*

4.2.2 Datasets

Sentinel-2 Data

Sentinel-2 mission [Drusch et al., 2012] is composed of two satellites, Sentinel-2A and 2B respectively launched in June 2015 and in March 2017. They have a high revisit frequency

Table 4.1 – *Sentinel-2 bands available with 2A product from Theia/Muscate.*

Band name	Central Wavelength (nm)	Spatial Resolution (m)
Band 2 - Blue	492.4	10
Band 3 - Green	559.8	10
Band 4 - Red	664.6	10
Band 5 - Vegetation Red Edge	704.1	20
Band 6 - Vegetation Red Edge	740.5	20
Band 7 - Vegetation Red Edge	782.8	20
Band 8 - Near Infra-Red	832.8	10
Band 8A - Vegetation Red Edge	864.7	20
Band 11 - SWIR	1613.7	20
Band 12 - SWIR	2202.4	20

of 5 days over the equator, and 2/3 days near mid-latitudes which is important to map land cover dynamics. Each sensor owns 13 spectral bands with different wavelengths, from the visible to the shortwave infrared at different spatial resolutions. For this research paper, satellite data used come from the Theia/Muscate database (`https://www.theia-land.fr/`) and 10 spectral bands are available for each sensor (Table 4.2.2). This product is available on their dissemination platform or by automatic download by requesting their server. For this work, cloudless single-date Sentinel-2 images, from 24th July 2019 and 21th July 2020, are respectively chosen for tiles 32ULU and 31UGQ.

UF Typology and Reference Dataset

From High Spatial Resolution Imagery (10m), many research papers map UF in western cities in 4 classes (Table 4.2.2): dense (class 1) and sparse built-up (class 2) areas where the main difference consists in their relative density and the importance of vegetated areas and bare surfaces, specialized areas (class 3) characterizing industrial activities or waste land or open areas with a majority of artificial or bare surfaces and large scale road or rail network (class 5). This is the case for the well-known product, OSO (Occupation des Sols Opérationnelle - `http://osr-cesbio.ups-tlse.fr/oso/`) available at the national scale and produced from times series of Sentinel-2 or other research works [El Mendili et al., 2020]. In order to improve this classification of UF, a fifth class has been added and describes specialized areas where green surfaces are dominant (more than 80%) (class 4) such as urban parks, cemeteries or vegetated sports or leisure complexes (outdoor sports

fields). In this paper, we assume that a spatial resolution of 10m make it possible to map these five UF classes. A sixth class for non-urban areas (class 6) corresponds to any other non built-up areas as agricultural lands, forests or water surfaces.

To produce this reference datastet adapted to any type of city in France to map UF at 10m spatial resolution, two existing topographic databases are used : (1) the regional landuse/cover vector database (BDOCSGE2©GeoGrandEst[3], 2019) and (2) the national topographic database with networks (BDTOPO©IGN[4], 2014). The Minimum Mapping Unit (MMU) is less than $50m^2$ for buildings and urban areas. The first database is produced by visual interpretation of aerial images (2018-2019) and maps each department of the region into 53 classes at level 4. The thematic classes are closer to landuse than landcover classes. This database is a freely available at large scale (1:10,000) from the GeoGrandEst Data Infrastructure (www.geograndest.org) and the production is ongoing for the whole region. At the time of this research, only five departments were available (Bas-Rhin, Moselle, Haut-Rhin, Vosges and Meurthe-et-Moselle) in relation with our tests sites. Since the satellite images and reference data are less than two years distant, we make the hypothesis that changes are minor due to the low population dynamics in the region. The legend of BDOCSGE2 is organized into four levels of nomenclature where the first level categorizes land cover into four classes (1) artificial surfaces, (2) agricultural areas, (3) forest areas, and (4) water surfaces. Artificial surfaces (Level 1) are further sub-divided into 16 classes at Level 3 and in 29 classes in level 4. In the BDOCSGE2 layer, all the roads have the same degree of importance which makes it impossible to remove the smallest polygons in the inner city that cannot be distinguished at 10 m spatial resolution (Table 4.2.2). In order to produce an adapted road network thematic class, the second database (BDTopo), produced by IGN describing lines in vector format with the degree of importance, is pre-processed. A buffer of 30 meters for the highways and 10 meters for the major roads is calculated to correspond to the real size of these networks which the only ones to be visible at 10 m. These data are then added to other classes of the vector reference datasets layer. We then summarize the fifth classes into a unique vector layer and rasterize alls polygons at a 10 m spatial resolution. The UF classes are not evenly distributed on the reference data and represent 8.4% of the total area. Indeed, Dense Built-Up represents about 8.4% of the total UF area of the dataset, Sparse Built-Up 59.5%, Specialized Built-Up Areas 19%, Specialized but Vegetative Areas 8.3% and Large Scale Networks 4.8%. For the last test with four urban fabric classes, the class (4) describing specialized but vegetative areas has been removed and all these areas have been included in the last class (6).

---

[3]https://www.datagrandest.fr/portail/fr/projets/occupation-du-sol
[4]https://geoservices.ign.fr/documentation/donnees/vecteur/bdtopo
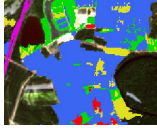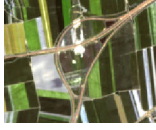
Table 4.2 – *List of the five UF classes (after pre-processing) used for this research (1/6000 scale).*

Class	Subset on Sentinel-2	Reference data	Description
(1) Dense Built-Up			Surfaces mainly occupied by buildings and impervious surfaces. Vegetation and bare soil are scarce.
(2) Sparse Built-Up			Buildings and other artificial surfaces share the land with green surfaces and bare soil
(3) Specialized Built-Up Areas			Surfaces allocated to production, commercial, service and tertiary activities
(4) Specialized but Vegetative Areas			Surfaces containing at least 80% of vegetated areas and 20% of bare soil or impervious surfaces as urban park, sport leisure activities, cemetery or campgrounds
(5) Large Scale Networks			Primary road network and others associated areas, railways and train stations
(6) Others non-urban areas			Every non-urban areas such as agricultural land, forests, wetlands and water surfaces

4.2.3 Methods

The proposed workflow is proposed in three different steps (Figure 4.2.: (1) the pre-processing step for a training, validation and test patches data preparation, (2) the model training step where four different approaches built the same base network are compared and (3) the post-processing and evaluation step to predict and test every approach. All models have been trained and tested on a computer with an RTX Quadro 4000 with 8 Gb of VRAM, an Intel(R) Xeon(R) E-2246G processor clocked at 3.6 GHz for 6 physical cores

53

and 32 Gb of rams. For implementation, we used Keras API [Chollet et al., 2015] built on top of TensorFlow 2.0.
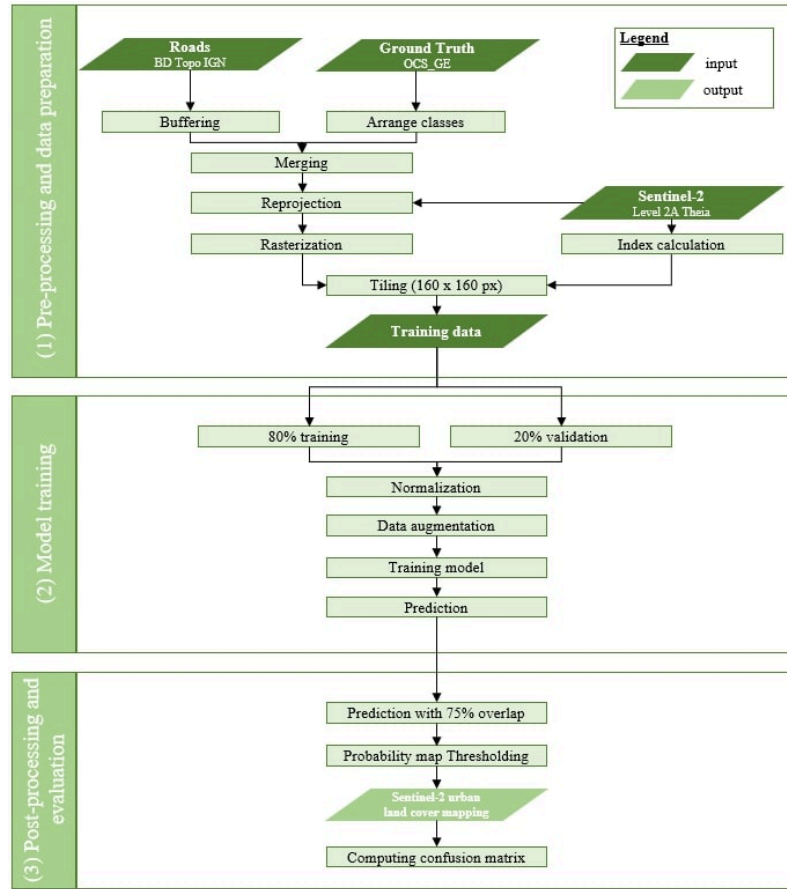
Figure 4.2 – *Workflow for automated UF mapping in Sentinel-2 L2A imagery over one tile. This workflow contains (1) preprocessing and data preparation where reference data, roads and satellite data are preprocessed, (2) model training where some networks are applied and (3) post-processing and evaluation where predictions are made.*

Step 1: Pre-processing and data preparation

On the Sentinel-2 images, first combination of spectral bands, as input for all the Networks, are the 3 IRRG bands (Infrared, red and green). We kept these three spectral bands to merge the two networks with a similar depth and compare the 4 methods developed with equivalent input data. In addition, band 8 (Near Infra-Red) was preferred to band 2 (Blue) because it provides more information on vegetation and allows better distinction between urban and natural areas. The normalized difference vegetation index (NDVI) [Rouse, 1973] (Equation 4.1), the normalized difference built-up index (NDBI) [Zha et al., 2003] (Equation 4.2) and the entropy [Haralick et al., 1973] based on NDVI index are also calculated and will be inputs in Networks due to their relevance in urban studies [Huang et al., 2019, Mhangara et Odindi, 2012, Su et al., 2008].

The reference data is pre-process in order to produce training and validation dataset for the different Networks used.

$$NDVI = \frac{Red - NIR}{Red + NIR} \qquad (4.1)$$

$$NDBI = \frac{SWIR1 - NIR}{SWIR1 + NIR} \qquad (4.2)$$

where *Red* is the red band of the Sentinel-2 image (band number 4), *NIR* is the near infrared band (band number 8) and *SWIR1* is the Short-Wave Infrared 1 band (band number 11). SWIR1 is resampled at 10m spatial resolution to match spatial resolution with other spectral bands.

Haralick et al. [1973] introduced the concept of Grey Level Co-occurence Matrix (GLCM) in 1973. This technique of feature extraction is widely used in the field of image analysis. GLCM represents a histogram of co-occurring greyscale values at a given offset. Then in urban areas where the heterogeneity is high, the Entropy index representing the randomness of disorder present in the image is calculated. The entropy value is high when the elements of the co-occurrence matrix are the same and lower are the values of entropy and more unequal are the elements. Entropy (eNDVI) is calculated using NVDI. We defined a direction of 3 pixels and an offset of 1 to compute the GLCM.

$$eNDVI = -\sum_i \sum_j p_d(i,j) \ln p_d(i,j) \qquad (4.3)$$

where $p_d(i, j)$ is the *(i, j)*th element of the normalized GLCM.

Data sources are split in training, validation and test zones following Saraiva et al. [2020] methodology which consists in selecting an area covered by the reference data, splitting it with a part for the training set and another for the validation set. The patches of each set are then cut in the selected areas by applying an overlap of 50% between each patch. Each area being independent, there can be no overlap between the partitions. Random patches are selected in training and validation areas which result of 5,517 training patches (80%) and 1,379 validation patches (20%) for 31UGQ tile and 8,942 training patches (80%) and 2,685 validation patches (20%) for 32ULU tile (Table 4.2.3). Test zone is splited in patches with 75% overlap and reconstructed to predict and evaluate all the test zone. This method is applied for every network tested.

Table 4.3 – *Number of patch containing a class and number of pixels per class for each training set.*

Class	32ULU		31UGQ	
	Patches	Pixels	Patches	Pixels
(1) Dense Built-Up	3,825	1,295,473	2,330	849,901
(2) Sparse Built-Up	7,154	10,724,491	4,374	4,698,560
(3) Specialized Built-Up Areas	5,050	3,141,083	4,168	1,985,748
(4) Specialized but Vegetative Areas	6,147	1,802,533	3,612	1,228,366
(5) Large Scale Networks	1,420	502,951	1,148	477,968
(6) Others non-urban areas	8,942	211,448,669	5,517	131,994,657

Table 4.4 – *List of executions including networks used and spectral bands/index.*

Method	Network	Spectral bands/Index
U-Net-IRRG (Figure 4.4)	U-Net pretrained on ImageNet	3, 4, 8
U-Net-Index (Figure 4.4)	U-Net not pretrained	3, 4, 8, NDVI, NDBI, eNDVI
U-Net-Encoder (Figure 4.5)	Fusion of two U-Net	3, 4, 8 and NDVI, NDBI, eNDVI
U-Net-Decoder (Figure 4.6)	Fusion of two U-Net	3, 4, 8 and NDVI, NDBI, eNDVI

Models training

This section gives a detailed description of models training and selected networks. Fusion methods were successfully tested in our domain application for encoder/decoder like networks [Hazirbas et al., 2016, Audebert et al., 2018, Jie et al., 2020] to combine different input data sources for land cover classification.

In this paper, one main U-Net network with VGG-16 is used as the encoder to be able to apply pretrained ImageNet [Russakovsky et al., 2015] weights. We chose this backbone because it is one of the smallest networks and helps to limit overfitting. Pretrained weights and backbone VGG-16 network were obtained from keras built-in models. The encoding phase for this network allows to extract a set of features to detect the classes present within the patches. Conversely, the decoding phase allows to restore the spatial context of the patch. Skip connections between the two phases of the U-Net network allows to find more quickly characteristics discovered during the first blocks of the network without having to go through the deeper meshes.

We have developed two fusion methods inspired by Hazirbas et al. [2016] works as it has proven to be effective in Audebert et al. [2018] by combining exogenous indexes with reflectance data. First, a fusion of the encoders of two U-Net is performed (Figure 4.5)

and then a fusion of the decoders (Figure 4.6). These fusion methods are compared to the classical U-Net network (Figure 4.4) with different input parameters described below and summarized in Table 4.2.3:

- `U-Net-IRRG (Figure 4.4):` U-Net with VGG-16 as a backbone, taking IRRG patches as inputs. This network is pretrained on ImageNet ;

- `U-Net-Index (Figure 4.4):` U-Net with VGG-16 as a backbone, taking six channels patches as inputs : green, red, infrared bands and NDVI, NDBI and eNDVI. This network could not be pretrained on ImageNet because it takes more than 3 channels as input ;

- `U-Net-Encoder (Figure 4.5):` Two U-Net with VGG-16 as as backbone. The main network takes 3 indexes patches (NDVI, NDBI and eNDVI) as inputs and it is not pretrained on ImageNet. The second network takes IRRG as inputs and is pretrained on ImageNet. The contributions of each encoder are summed after each convolution block ;

- `U-Net-Decoder (Figure 4.6):` The same methodology as encoder fusion is applied but the fusion is executed during the decoder phase, including bottleneck in order to alter the final result as much as possible ;

These approaches consist in classifying satellite image patches of dimension $h \times w \times n_{channels}$ to obtain a classification of dimensions $h \times w \times n_{classes}$. All the operations described in Figures 4.4, 4.5 and 4.6 are explained below.

- `Convolution:` Each convolution block consists in applying a convolution with a kernel size of 3×3 and a stride of 1 pixel. *ReLu* (Rectified Linear Unit, $f(x) = max(0, x)$) is the activation function used at the end of each convolution ;

- `Dropout:` This regularization technique is commonly used when developing a neural network [Srivastava et al., 2014]. 50% dropout [Pelletier et al., 2019] is applied between the two convolution blocks during the decoding phase to limit overfitting. It consists of a random and temporary deactivation of some neurons to avoid complex co-adaptation. During the prediction phase, neurons are reactivated to test the new model. We decided to use dropout layers because our training data are small and imbalanced. Also, dropout has been used in different work as it provides restrictive regularization and enhance generalization [Rajaraman et al., 2020] ;

- `MaxPooling:` The MaxPooling block allows you to reduce the size of the data during the encoding phase to detect different characteristics. For our network, we reduce the height and width of the features by 2 after all the convolution steps ;

- `Concatenation:` For this layer, the features of the same dimension $h \times w \times n$ of the decoding and encoding step are concatenated before the transpose step ;

- `Transpose:` For the decoding phase, a transpose layer was preferred to an Up-Sampling layer as it is commonly used in encoder/decoder architecture to perform semantic segmentation [Iglovikov et Shvets, 2018, Ouyang et Li, 2021]. Indeed, it is a more complex operation that combines a convolution operation and an upsampling operation within the same layer ;
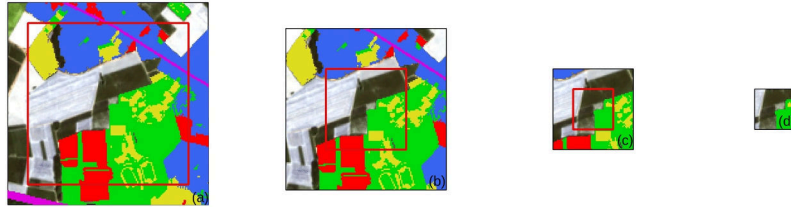
Figure 4.3 – *Size comparison of the patches: (a) 160x160 pixels, (b) 128x128 pixels, (c) 64x64 pixels and (d) 32x32 pixels.*

The size of the input patches are $160 \times 160 \times 3$ for each of the two networks used, with 5 land use classes at the output. This size allows the network to get a wider spatial context of the land use classes represented in the image (Figure 4.3). This size allows for a large footprint and diversity of classes on each patch. At the end of the network, a $\vec{x}$ vector of size 160×160 containing the semantic segmentation into 5 classes is produced. It is then normalized using a Softmax function defined below :

$$f(\vec{x})_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{4.4}$$

where i and j represent respectively the ith and j^{th} class and x_i the probability of belonging to the i class.
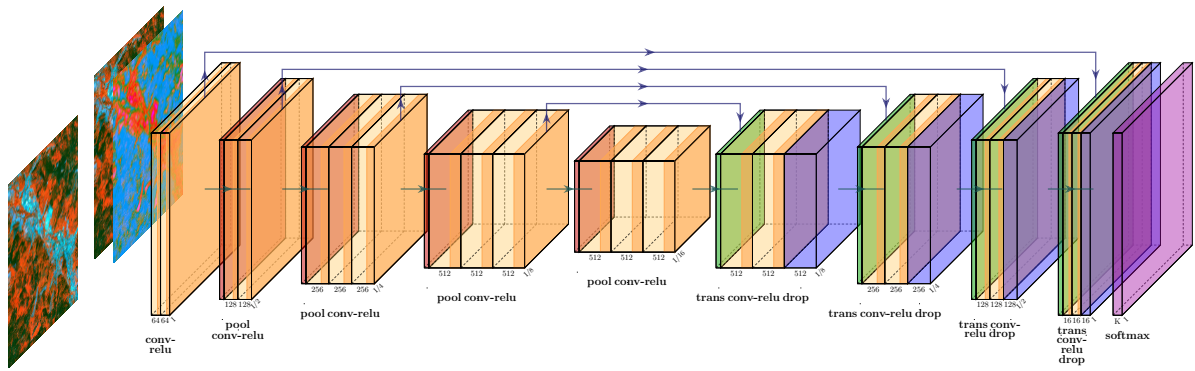
Figure 4.4 – *U-Net-IRRG and U-Net-Index architecture used for preliminary tests. This network uses, on the one hand, only IRRG images and, on the other hand, a combination of IRRG images and spectral and textural indexes (NDVI, NDBI and eNDVI).*

Several data augmentation methods were applied to the training patches to enrich the dataset. Each patch is thus kept in its initial state and then augmented randomly either by rotations (90, 180, and 270 degrees) or flipping (from left to right or from top to bottom) using *numpy* library[5]. These data augmentation methods doubled the size of the initial training sets of each tile. We also add Dropout and L2 regularization during decoder process [Srivastava et al., 2014] to reduce the chance of overfitting due to the use of an imbalanced dataset and a low number of patches. L2 was prefered to L1 as it more stable by putting half of the weights on each input [Li et al., 2021].

---

[5]https://numpy.org/
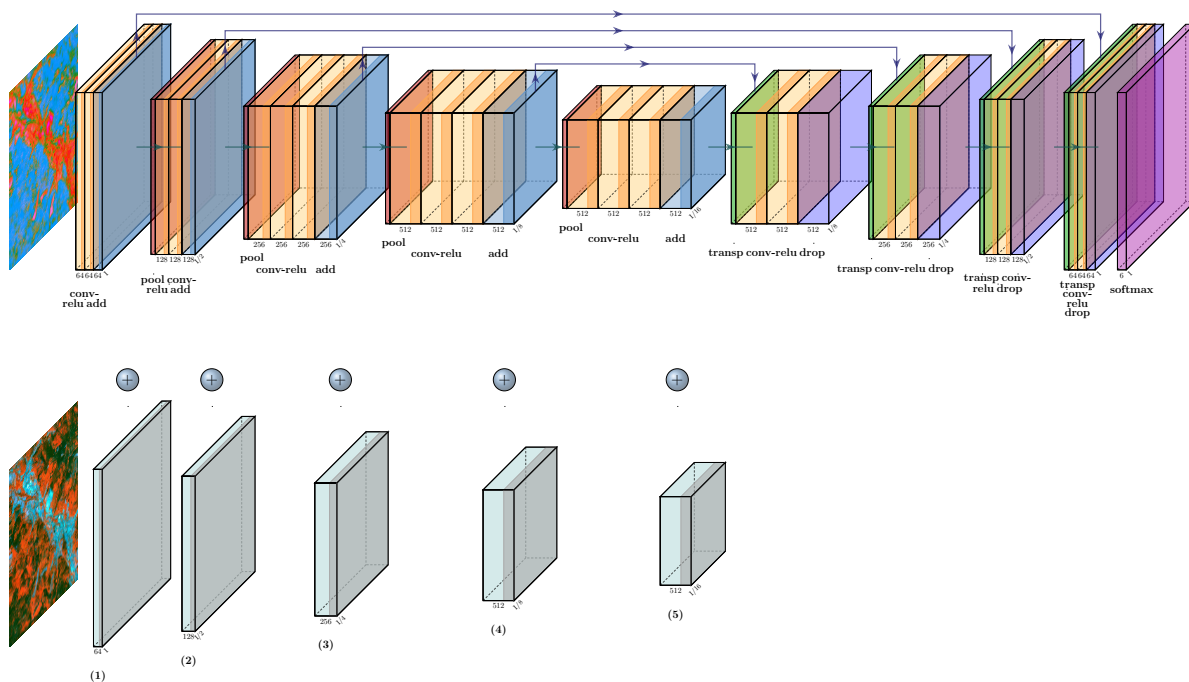
Figure 4.5 – *U-Net-Encoder architecture modified in order to apply encoder fusion. Features (1) to (5) are extracted from the second U-Net, pretrained on ImageNet with IRRG patches as inputs, and merged in the first U-Net during the encoding phase.*
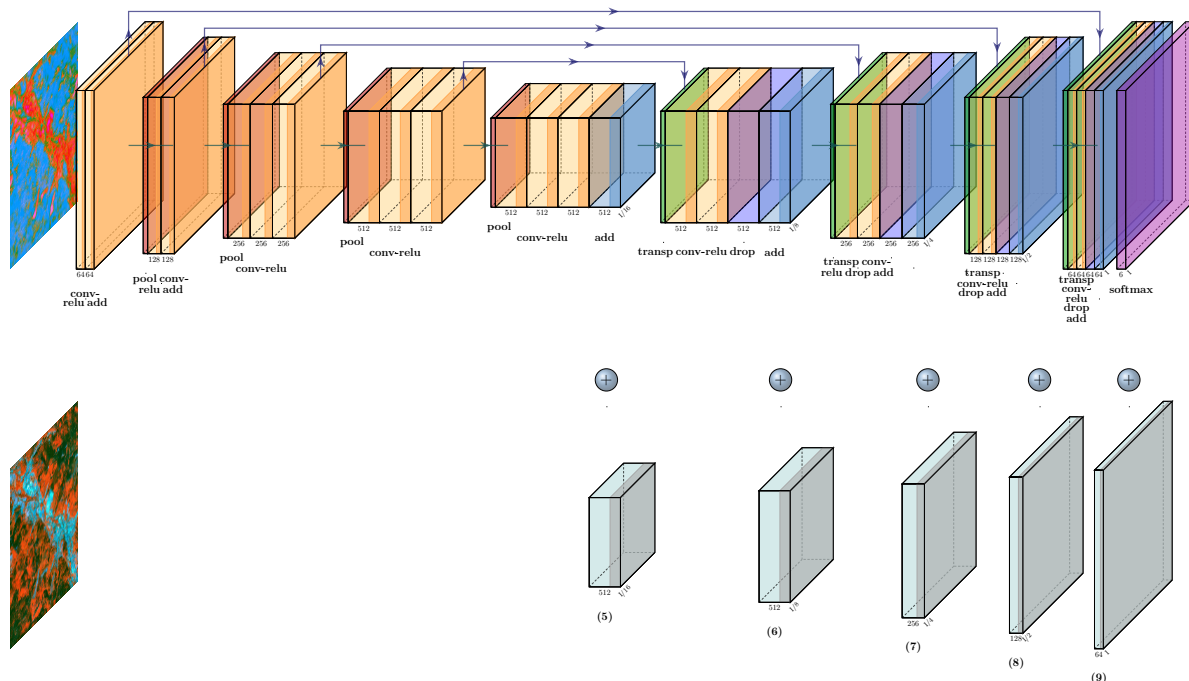
Figure 4.6 – *U-Net-Decoder architecture modified in order to apply decoder fusion. Features (5) to (9) are extracted from the second U-Net, pretrained on ImageNet with IRRG patches as inputs, and merged in the first U-Net during the decoding phase.*

Every image is normalized by dividing the standard deviation of the reflectance of the spectral band by the difference of this one with the average of the reflectances. This

normalization method allows the data to be centered on the same range of values so that our gradient remains stable. The normalization formula is developed below:

$$n = \frac{(b - \bar{b})}{\sigma_b} \tag{4.5}$$

where n represents the normalized spectral band, b the reflectance values of the spectral band, $\bar{b}$ the mean of the reflectance values, and σ_b the standard deviation of the reflectance values.

The models were trained for 100 epochs with a Learning Rate (LR) of 1×10^{-4} and a batch size of 8. We reduced LR by 50% each time a plateau was reached for 5 epochs [Chhor et Aramburu, 2017] using Keras callback ReduceLROnPlateau. Softmax was used as an activation function for the last layer of each model to predict multinomial probabilities (Figures 4.4, 4.5 and 4.6).

Training takes around 15 hours for each network.

Post-processing and evaluation

After training the model, predictions are made on selected test areas. The test image must first be reconstructed from 160 x 160-pixel patches. To do this, an overlap of 75% is applied to smooth the predictions and improve the classification results. The overlapped pixels within the overlap regions are averaged and then the index of the band with the highest probability is retained as a prediction. This method is applied to all the test images. Finally, a confusion matrix is computed over the entire image accompanied by a file with detailed statistics by land-use class. This prediction technique can also be applied to the whole image to provide an accurate mapping of land use in urban structures.

4.2.4 Loss function

To take into account the low representativeness of certain classes in the dataset (imbalanced dataset), a weighted categorical cross-entropy loss is used by assigning higher weights to the classes with the least surface area. These are the inverse of the class frequency [Audebert et al., 2018]. This loss is commonly used in remote sensing multi-class supervised classification tasks [Ienco et al., 2017, Zhu et al., 2017b]. The loss has been taken from *segmentation_models* framework [Yakubovskiy, 2019]. On the other hand, the "Large Scale Networks" and "Specialized But Vegetative Areas" classes (rare objects composing the urban structures) occupy such small areas that we have decided to assign the lowest weights to another poorly represented "Dense Built-Up" class. Indeed, using too high weights can

bias the loss and distort the learning process which would lead to prediction errors [Sefrin et al., 2021, Audebert et al., 2018].

4.2.5 Evaluation metrics

We adopted different evaluation metrics [Maxwell et al., 2021a] to measure the quality assessments and the effectiveness of every network tested in our processing chain : Precision, Recall and $F1_{Score}$.

Precision (also know as User's Accuracy - UA) informs about the fraction of well-classified pixels in the classified image. It can be calculated by dividing True Positives (TP) values with the sum of True Positives (TP) and False Positives (FP).

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

Recall (also know as Producer's Accuracy - PA) indicates the fraction of well-ranked pixels relative to the reference data. It can be calculated by dividing True Positives (TP) values with the sum of True Positives (TP) and False Negatives (FN).

$$Recall = \frac{TP}{TP + FN} \tag{4.7}$$

$F1_{Score}$ (also known as Dice) represents the harmonic mean between Precision and Recall. It is calculated by dividing twice the product of Precision and Recall by the sum of these same metrics. In order to have a global analysis metric of the results, the $F1_{Score}$ weighted is calculated for all the classes for each test.

$$F1_{Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.8}$$

4.3 Results

In order to assess the results of the four networks the section 3.1 is dedicated to a global analysis comparing the test sites without distinction of UF. The section 3.2 presents the results for each study areas (for remind, three test sites with a gradient of urbanisation, respectively Strasbourg, Metz and St-Avold city - Figure 4.1). Then, section 3.3 focuses on the best fusion methods considering qualitative and quantitative results and the last section (3.4) compared this latest with a test mapping UF in four classes.

4.3.1 Global results analysis

The weighted $F1_{Score}$ and Overall Accuracy has been calculated for each test area (Tables 4.5 and 4.6). We notice a slight advantage for the U-Net-Decoder method for the two least dense cities, Metz and Saint Avold (respectively 0.7488 and 0.7883 for weighted $F1_{Score}$ and 0.7343 and 0.7580 for Overall Accuracy). The U-Net-Encoder method obtains the high value for the city of Strasbourg with a weighted $F1_{Score}$ of 0.5990 and an Overall Accuracy of 0.5794. We can notice that the difference is however really close between U-Net-Encoder/Decoder. A detailed analysis for each UF is then necessary to conclude on their performance to map UF whatever the thematic classes.

U-Net-IRRG			U-Net-Index		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5133	0.7364	0.7716	0.5005	0.7214	0.7248
U-Net-Encoder			U-Net-Decoder		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5990	0.7479	0.7834	0.5894	**0.7488**	**0.7883**

Table 4.5 – *Results of weighted F1$_{Score}$ for each method in every study area.*

U-Net-IRRG			U-Net-Index		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5004	0.7067	0.7244	0.4737	0.6963	0.6701
U-Net-Encoder			U-Net-Decoder		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5794	0.7299	0.7513	0.5706	**0.7343**	**0.7580**

Table 4.6 – *Results of Overall Accuracy for each method in every study area.*

Training and validation loss has been plotted in order to monitor any possible overfitting (Figure 4.7).
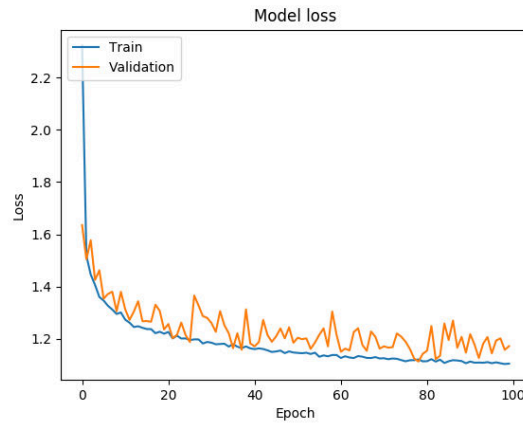
Figure 4.7 – *Training and validation learning curve for U-Net-IRRG method perform on 32ULU tile.*

4.3.2 Results analysis for each UF

A more detailed analysis of the evaluation metrics is presented from the most important city (Strasbourg) to the most rural test site (St-Avold). For each test area, quantitative analysis based on Precision, Recall and $F1_{Score}$ measures is completed by a qualitative analysis of results with some zooms on the three test areas.

Semantic segmentation results for Strasbourg, Grand Est, 32ULU Tile

Table 4.7 summarizes the three evaluation metrics for the four different methods and for each UF (Table 4.2.2) based on Strasbourg test area. We notice an advantage for both fusion methods (U-Net-Encoder and U-Net-Decoder) where the best statistical results are found for both methods. More precisely, these fusion methods improve the results of the specialized but vegetated area (Class 4) and the large scale networks (Class 5), where the $F1_{Score}$ is respectively 0.4716 and 0.5038 for U-Net-Encoder and 0.4716 and 0.5781 for U-Net-Decoder.

	U-Net-IRRG			U-Net-Index		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Class 1	0.3928	0.6820	0.4985	**0.5541**	0.3929	0.4598
Class 2	0.6116	0.3752	0.4651	0.7412	0.2832	0.4098
Class 3	0.5045	0.5789	0.5391	0.3749	**0.7388**	0.4974
Class 4	0.3279	0.3313	0.3296	0.3988	0.3236	0.3573
Class 5	0.2887	0.7543	0.4176	0.2298	**0.8325**	0.3602
Class 6	0.9876	0.5199	0.6812	0.9798	0.5853	0.7328
	U-Net-Encoder			U-Net-Decoder		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Class 1	0.4511	**0.6836**	0.5435	0.4797	0.6297	**0.5446**
Class 2	0.7415	**0.4651**	**0.5717**	**0.7531**	0.4179	0.5375
Class 3	**0.5169**	0.5878	0.5501	0.4773	0.7006	**0.5678**
Class 4	**0.4057**	0.5632	**0.4716**	0.3801	**0.6210**	**0.4716**
Class 5	0.3962	0.6917	0.5038	**0.5099**	0.6673	**0.5781**
Class 6	**0.9882**	**0.6427**	**0.7789**	0.9880	0.5987	0.7456

Table 4.7 – *Results of all methods for the test zone located in Strasbourg, Grand-Est, France.*

To complete these quantitive results, Figure 4.8 presents some subsets in the Strasbourg test area. Compared to the reference dataset, qualitative analysis shows that the methods U-Net-Encoder (subfigure e) and U-Net-Decoder (subfigure f) provide a better detection of large scale networks (class 5) than the reference data (Figure 4.8 b). Indeed, some roads and railway are classified while they do not appear on the image (Figure 4.8 b). The specialized built-up areas (Class 3) is also well extracted by all methods by detecting areas with ongoing construction that are also not present on the reference data. On the other hand, we notice an overestimation of the Specialized but vegetative areas (class 4) which includes some cropping or forests areas in peri-urban area. U-Net-IRRG and U-Net-Index also produce some confusion between different classes, such as Specialized Built-up Areas (Class 3) and Specialized but vegetative Areas (Class 4), where the two fusion methods have much more unified results.
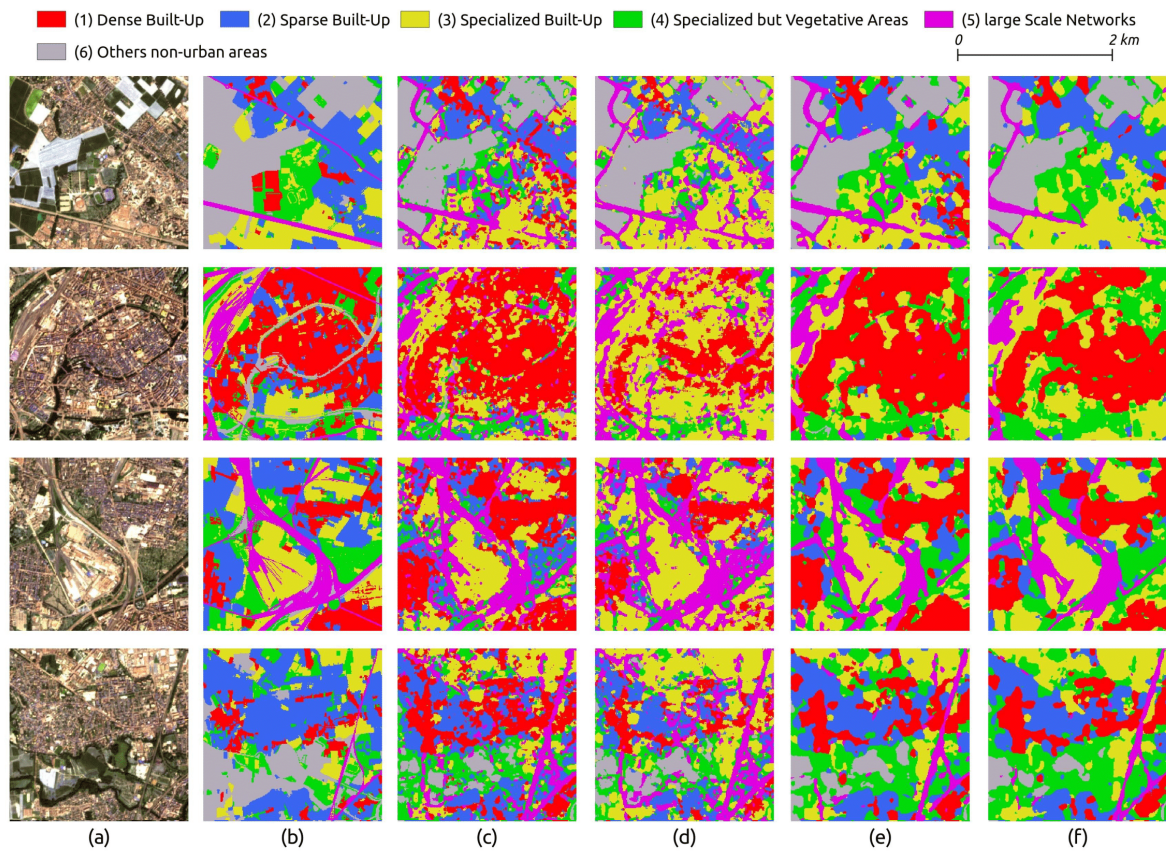
Figure 4.8 – *Semantic segmentation results for test zone over Strasbourg. (a) and (b) represent subset image and reference data respectively, (c) and (d) are respectively U-Net-IRRG and U-Net-Index and (e) and (f) are U-Net-Encoder and U-Net-Decoder.*

Semantic segmentation Results for Metz, Grand Est, 31UGQ Tile

Table 4.8 summarized all the statistics calculated by class for each method for the second test area located in Metz and its surroundings.

	U-Net-IRRG			U-Net-Index		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	**0.5128**	**0.7021**	**0.5199**	0.4577	0.5095	0.4822
Class 2	**0.6228**	0.6941	**0.6565**	0.5286	0.6586	0.5864
Class 3	**0.6177**	0.4329	0.5090	0.5385	0.5359	0.5372
Class 4	0.2092	**0.6298**	**0.3141**	0.1662	0.2666	0.2048
Class 5	0.4457	0.8559	0.5862	0.2812	**0.9090**	0.4295
Class 6	**0.9643**	0.7678	0.8549	0.9595	0.7758	0.8579
	U-Net-Encoder			U-Net-Decoder		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.4510	0.5847	0.5092	0.3858	0.6592	0.4867
Class 2	0.6136	**0.6948**	0.6516	0.6195	0.5962	0.6076
Class 3	0.5228	**0.5789**	0.5494	0.5026	0.6186	**0.5546**
Class 4	0.2193	0.4168	0.2874	**0.2674**	0.3735	0.3117
Class 5	**0.4638**	0.8471	**0.5994**	0.4240	0.8791	0.5721
Class 6	0.9601	0.7933	0.8687	0.9587	**0.8126**	**0.8796**

Table 4.8 – *Results of all methods for the test zone located over Metz, Grand-Est, France.*

The statistical results (Table 4.8) shows that the $F1_{Score}$ scores values follow the same trends the results of test area 1 (Strasbourg - section 3.2.1), with class 6 (Others non-urban areas) always higher than the other classes (above 0.85) and classes 1, 2, 3 and 5 always equal or above 0.5. Large Scale network (Class 4) still has low F1 values, even slightly lower than test area 1. This can be explained by the urban morphology of the city characterized with a lot of sparse urban settlements. These trends are confirmed in the precision and recall values. They are not significantly different than in subset 1 but the maximum values are reached here by the U-Net-IRRG model in terms of precision and the U-Net-Encoder model for recall. More precisely, the U-Net-IRRG model underestimates classes 1 to 3 while the U-Net-Encoder model overestimates classes 2 and 3 (Sparse and Specialized Built-Up areas).

Qualitative results (Figure 4.9) confirm that U-Net-Encoder and Decoder show better extraction of all the classes. Large Scale Networks (Class 5) is also better detected and new roads not visible in reference data are extracted. The U-Net-IRRG method clearly overestimated class (3). Specialized but Vegetative Areas (class (4) are much better detected visually for both fusion methods than for U-Net-IRRG.

Figure 4.9 – *Semantic segmentation results for test zone over Metz. (a) and (b) represent subset image and reference data respectively, (c) and (d) are respectively U-Net-IRRG and U-Net-Index and (e) and (f) are U-Net-Encoder and U-Net-Decoder.*

Semantic segmentation results for Saint-Avold, Grand Est, 31UGQ Tile

For this last test site (Table 4.10), Saint-Avold, Grand Est, the trends of the statistical results are identical or even with lower $F1_{Score}$ values particularly for the specialized but Vegetated Areas which is the most complex class. This means that results are lower when the urbanisation gradient decrease, more sparse are urban settlements and more difficult is the extraction.

	U-Net-IRRG			U-Net-Index		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Class 1	0.2597	**0.8114**	0.3934	**0.3209**	0.5785	**0.4128**
Class 2	0.6233	0.7010	0.6599	0.5202	0.6525	0.5788
Class 3	**0.3902**	0.4534	0.4194	0.3379	0.5544	**0.4199**
Class 4	0.0838	**0.3414**	0.1346	0.0348	0.1378	0.0556
Class 5	0.4571	0.9194	**0.6106**	0.3364	**0.9558**	0.4977
Class 6	**0.9944**	0.7493	0.8546	0.9896	0.6925	0.8148
	U-Net-Encoder			U-Net-Decoder		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Class 1	0.2996	0.5668	0.3920	0.2593	0.6958	0.3778
Class 2	**0.6367**	**0.7153**	**0.6737**	0.6359	0.6114	0.6234
Class 3	0.2306	0.6118	0.3350	0.2350	**0.6462**	0.3446
Class 4	0.1186	0.2455	0.1600	**0.1566**	0.2320	**0.1870**
Class 5	**0.4633**	0.8944	0.6104	0.4363	0.9091	0.5896
Class 6	0.9933	0.7765	0.8716	0.9930	**0.8034**	**0.8882**

Table 4.9 – *Results of all methods for the test area located near Saint-Avold, Grand-Est, France.*

The qualitative analysis (Figure 4.9) confirm the quantitative results by showing that extraction of all classes are overestimated for the first two methods (U-Net-IRRG and U-Net-Index). As in the other both test areas, class 5 stays better than the reference data by considering the major part of the large scale networks. Moreover, the fusion methods (U-Net-Encoder and U-Net-Decoder) propose a much more unified and smoothed classification results than the two others methods which is also explained with the recall results.
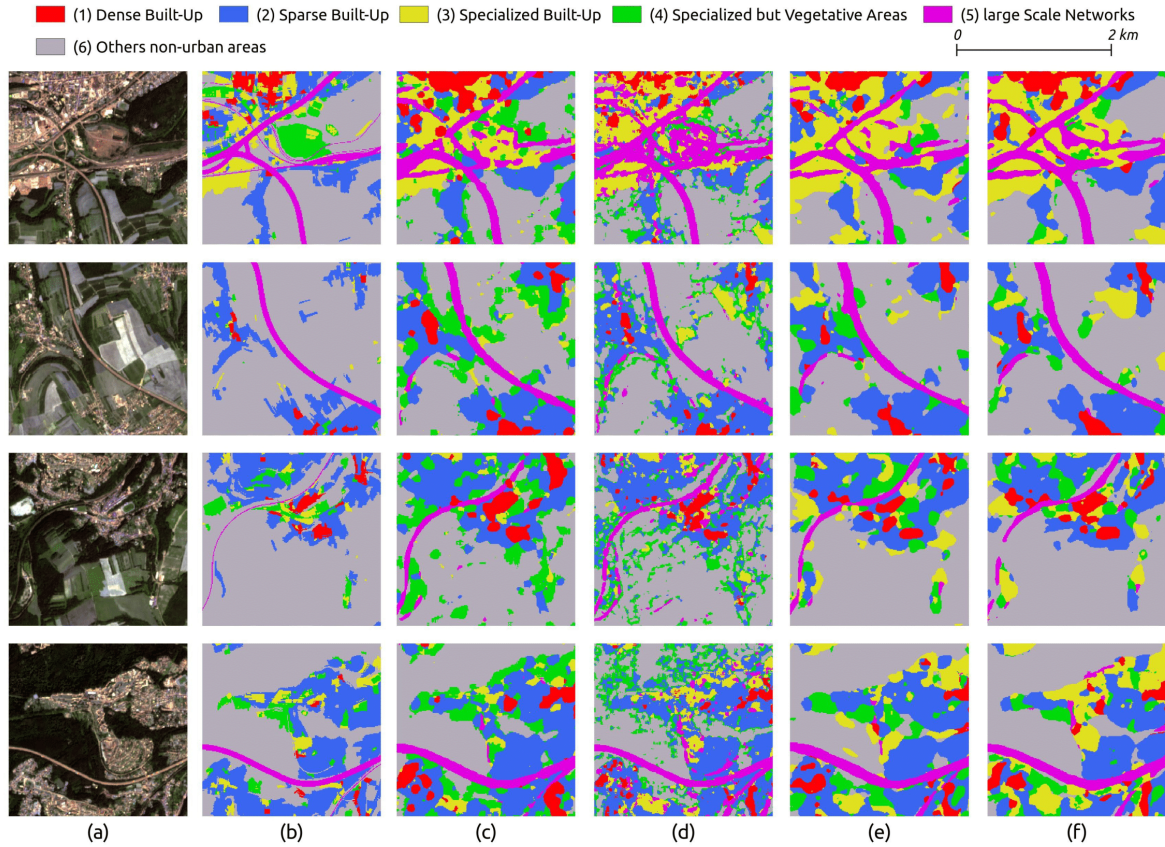
Figure 4.10 – *Semantic segmentation results for test zone over Saint-Avold. (a) and (b) represent subset image and reference data respectively, (c) and (d) are respectively U-Net-IRRG and U-Net-Index and (e) and (f) are U-Net-Encoder and U-Net-Decoder.*

In order to identify which of the both fusion methods is better to extract these 5 classes related to the test areas chosen for their urbanisation gradient, a more detailed analysis is proposed in the next section.

4.3.3 Encoder and Decoder fusion analysis

This section presents a detailed analysis based on 2D scatter plots for each metrics where the six classes are located in the 2D-space where the X-axis is the U-Net-Encoder and where the Y-axis is the U-Net-Decoder. Based on this Figure 4.11, we notice that, for all the metrics studied, UF classes are close to the diagonal line, which means that the both methods give close statistical results. Values of each metrics are concentrated around 0.5 with a Recall always slightly higher than $F1_{Score}$ with the U-Net-Decoder. Only the precision values are similar for the classes 1,2,3,5 and confirm that the class 4 is very difficult to extract but the U-Net-Decoder is the model for which all metrics are higher (Figure 4.13).
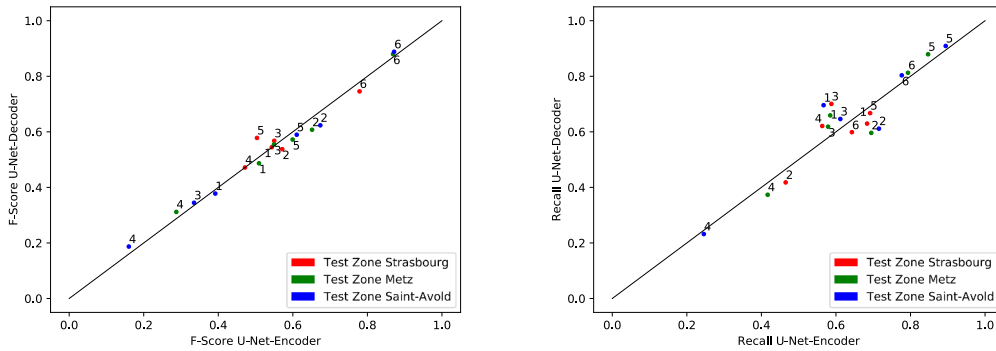
(a) $F1_{Score}$ for each class/test zones of the two fusion methods (U-Net-Encoder and U-Net-Decoder)

(b) Recall for each class/test zones of the two fusion methods (U-Net-Encoder and U-Net-Decoder)
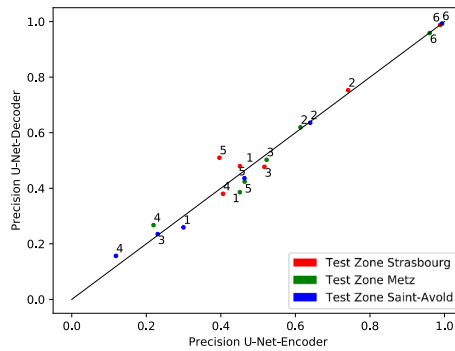
(c) Precision for each class/test zones of the two fusion methods (U-Net-Encoder and U-Net-Decoder)

Figure 4.11 – $F1_{Score}$, Recall and Precision for the analysis of the two fusion methods at the three study sites.

Figure 4.12 allows for a more in-depth analysis of the two fusion methods (U-Net-Encoder and U-Net-Decoder). The subset (i) focuses on a road not mapped in the reference data because it is not considered as primary road. However, both models detect it due to his width. The subset (ii) highlights a better precision in the delimitation of the Specialized but Vegetative Areas (class 4) for the U-Net-Decoder method. This same observation also applies to the subsets (iii) and (iv) where the Dense Built-up (class 1) and Specialized Built-Up Areas (class 3) are better extracted for the U-Net-Decoder method.
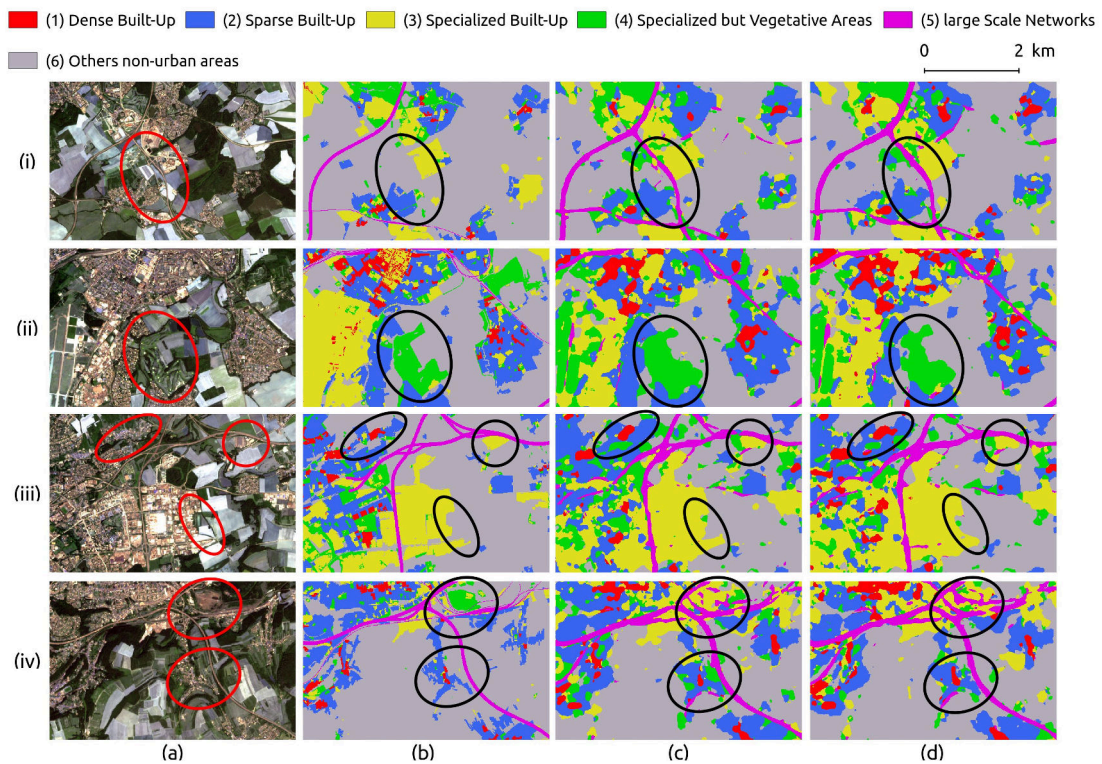
Figure 4.12 – *Four subsets are presented, from (1) to (4) with (a) Sentinel-2 zoom, (b) reference data, (c) U-Net-Encoder and (d) U-Net-Decoder.*
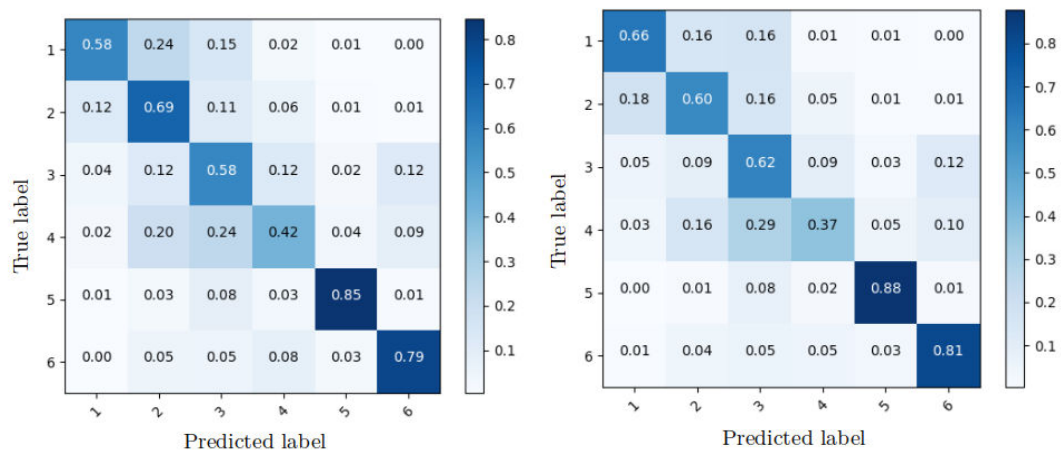
(a) Confusion matrix for U-Net-Encoder over Metz

(b) Confusion matrix for U-Net-Decoder over Metz

Figure 4.13 – *Confusion matrix (with Recall metric inside each cell) for U-Net-Encoder and U-Net-Decoder over the test area of Metz.*

4.3.4 Four-classes results for the Strasbourg study area

In order to analyse the impact of the number and choice of UF classes on the classification results, we haved tested the both models (U-Net-Encoder and U-Net-Decoder) with only

five classes by removing the most complex one at 10m spatial resolution (Specialized but Vegetative Areas - class 4) and grouping areas in the class 6 (Other). We notice that the results (Table 4.10) of $F1_{Score}$ are always higher than 0.5 for all classes only with U-Net-Encoder and U-Net-Decoder. Precision metrics shows that Large Scale Network (class 5) is always well detected and Specialized Built-Up Areas (class 3) is slighlty better extracted. Recall metrics confirm that Dense Built-Up (class 1) and Specialized Built-Up Areas (class 3) are more overestimated with U-Net-Encoder than U-Net-Decoder. These statistical results are also visible in (Figure 4.14). Finally, quantitative and qualitative interpretation results show the trends in results with very similar analysis with four or five UF classes that confirmed the slighted superiority of the U-Net-Decoder.

	U-Net-IRRG			U-Net-Index		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Class 1	0.4674	0.6066	0.5281	0.4032	0.5232	0.4554
Class 2	**0.6164**	0.3534	0.4492	0.5035	0.3080	0.3822
Class 3	0.3696	0.6941	0.4824	0.3992	0.7110	0.5113
Class 5	0.2867	0.8279	0.4259	0.2667	**0.8354**	0.4043
Class 6	0.9844	**0.5170**	**0.6779**	**0.9930**	0.4908	0.6569
	U-Net-Encoder			U-Net-Decoder		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Class 1	0.4426	**0.6752**	**0.5347**	**0.4897**	0.5544	0.5201
Class 2	0.5393	**0.5841**	0.5608	0.5629	0.5838	**0.5732**
Class 3	**0.4268**	0.6892	**0.5271**	0.3958	**0.7362**	0.5148
Class 5	**0.4998**	0.6276	**0.5565**	0.4585	0.6835	0.5488
Class 6	0.9907	0.4901	0.6558	0.9914	0.5060	0.6700

Table 4.10 – *Four UF classes results of all methods for the test area in Strasbourg, Grand-Est, France.*
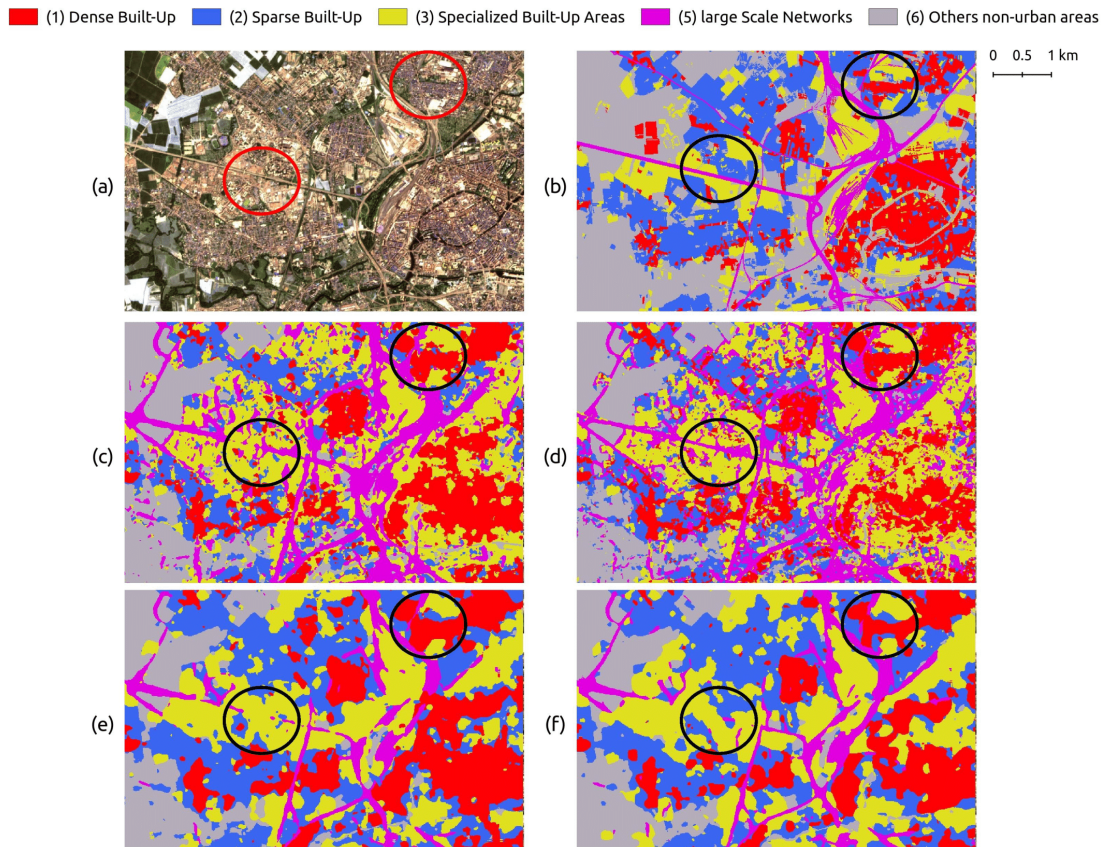
Figure 4.14 – *Semantic segmentation results for test area over Strasbourg in 4 UF classes with (a) S2, (b) reference data, (c) U-Net-IRRG, (d) U-Net-Index and respectively (e) and (f) with U-Net-Encoder and U-Net-Decoder.*

4.4 Discussion

The experiments performed in this study have tested two semantic segmentation networks (U-Net) that we combined using feature fusion between a from scratch network and a pre-trained network on ImageNet, to propose a generic UF mapping adapted to Grand-Est/France cities. First, the results showed the advantage of both fusion methods for the detection of these classes with quantitative and qualitative assessment showing better results for U-Net-Decoder method. Moreover, the contribution of exogenous indexes coupled with a pre-trained network allowed to refine the classification of the five UF classes.

Statistical and qualitative results confirm the good results of the U-Net-Decoder method over the U-Net-Encoder one to extract several UF classes with a mono-temporal image. It is possible to notice a better detection of some classes compared to the reference data. This is the case for Large Scale Networks (class 5) where roads and railroads, not yet or barely built at the time of the creation of the reference data, are well detected. Indeed, the selection of the Large Scale Networks was made according to a level of road in a French national database. Thus, there may be some roads of a lower level that are still

visible at 10m spatial resolution. But for most of the roads available at this level, they are not visible at this resolution. On the other hand, Specialized But Vegetative Areas (class 4) offer rather low scores rarely exceeding 0.40 of $F1_{Score}$. This is due to the complexity of this class at 10m where vegetated areas are dominant. The reflectance of these surfaces is quite similar to other vegetated areas such as crops, grasslands or small forests summarized in the Other non-urban areas (class 6). However, Recall values reach a value of more than 0.6 with the U-Net-Decoder model close to the other Recall values. This class (4) is also best detected when the areas is highly urbanized such as in the center of Strasbourg (test area 1) or Metz (test area 2). Indeed, green areas inside cities are very often belong to class (4) as they are mostly urban parks, leisure activities areas, campgrounds or cemetery. These surfaces are rather composed of green areas than mineral surfaces. They are most often detected as Specialised but Vegetative areas (class 4).

We also notice that the first methods (U-Net-IRRG) have many geometric errors in the delimitation of classes. In fact, the delimitation of class (4) has a significant number of overestimation. Each fusion method proposes a better estimation of these surfaces and can be seen with the $F1_{Score}$ metric. This is notably the case of U-Net-Decoder which takes them better into consideration by limiting the overestimation, as seen during the qualitative analysis. This is explained by the combination of intermediate features coming from the pre-trained network and bringing information in the delimitation of these surfaces. However, these surfaces remain complex to detect using mono-temporal imagery, even if the contribution of features from another network improves the qualitative and quantitative results. Moreover, the ImageNet weights have been initialized on RGB images. It could be interesting to test the impact of RGB bands compared to IRRG bands when using a pre-trained network.

Indeed at 10m, confusions apprears between Dense Built-up (1) and Sparse Built-Up (2) classes due to the low distance between buildings not visible at this resolution. Others confusions exist between Specialized but Vegetative Areas (4) and Others non-urban areas (6) due to the presence of vegetation on the urban frange. Despite the low number of Large Scale Networks patches (5) (Table 4.2.3), this class obtains encouraging scores and the networks are also able to classify roads that do not initially appear in the reference data (Figure 4.9). Overall, results remain encouraging despite the use of mono-temporal imagery. Nevertheless, the addition of multi-temporal and multi-source data should improve the current results even if urban environments have a low intra-temporal variability compared to natural areas.

The results from the both methods U-Net-IRRG and U-Net-Index give more hetero-

geneous classification results. The contribution of the feature fusion allows to obtain smoother UF classes with a better delimitation between the classes. This can be explained by the contribution of feature of a network taking in input only spectral and textural indices which allow to have a more precise information on the delimitation of urban classes. On the other hand, the network using spectral bands allows to extract more features allowing the distinction between the different UF classes. The U-Net has been designed to learn its own spatial filters. Thus, the contribution of eNDVI in conducted methods could be explored since the entropy is nothing else than a set of convolutions on a grayscale image with a sliding window.

Even with four classical classes rather than five UF classes, the results remain similar to the previous ones. This shows the interest of studying UFs as 5 classes rather than 4 because the Specialized but Vegetative Areas class is needed by end-users and is an integral part of urban areas.

4.5 Conclusions and perspectives

The objective of this research was to show the interest of semantic segmentation methods to help end-users to produce a relevant and up-to-date map of UF in five thematic classes only with an optical mono-temporal and high spatial resolution image (Sentinel-2). Indeed, classically with a high spatial resolution (10m), UF are only mapped with four classes distinguishing network from dense, sparse built-up areas and activities. In this paper, two methods using features fusion techniques (U-Net-Encoder and U-Net-Decoder) between two networks have been developed using (i) three spectral bands (green, red, NIR) for the pre-trained network and (ii) three spectral and textural indexes (NDBI, NDVI and *eNDVI*) for the non-pre-trained network. These methods were compared with a U-Net taking into account either IRRG or IRRG bands and three spectral and textural indexes. The idea was to combine two types of data, each providing various information to improve the detection of urban surfaces proposed in five different UF classes adapted to Grand-Est/France cities: (1) Dense Built-Up, (2) Sparse Built-Up, (3) Specialized Built-Up Areas, (4) Specialized but Vegetative Areas, (5) Large Scale Networks. Methods have been tested on a gradient of urban areas in the East of France to ensure a generalisation of results.

Those highlight that both fusion methods, especially the U-Net-Decoder one for most of UF, offer the best results in refining the detection of the most UF classes. The U-Net-Decoder method showed an advantage in the delimitation of Specialized but Vegetative Areas and a better classification of these areas for highly urbanized areas (Strasbourg

center and Metz). The qualitative analysis confirm these first analysis by showing an advantage for the U-Net-Decoder method with a better segmentation of the different UFs. On the other hand, the statistical results showed a better but close classification of the Dense Buit-Up and Sparse Built-Up for the U-Net-Encoder method. Most of UF classes are always extracted with relevant evaluation metrics (greater than 0.5) and a qualitative interpretation of the results shows homogeneous extraction patches of classes. The weakness of the results for the Class 4 which is confused with other land cover classes due its relative complexity, could be improved by using multi-temporal imagery in order to take into account the vegetation dynamics of this class. These first relevant results could be also applied to other study sites in France to confirm our conclusions.

This opens several perspectives for the use of optical times series imagery in order to take into account the spatio-temporal dynamic of UF. An other issue would be to integrate the multivariate properties of UF from Sentinel-1 imagery which has already demonstrated its interest for mapping urban footprint. Indeed, many works show the interest of using the radar amplitude for the detection of several UF classes. Our next research will focus on the addition of these multi-temporal optical and radar data for the mapping of these UF classes.

Multimodal and multitemporal land use/land cover semantic segmentation on Sentinel-1 and Sentinel-2 imagery : an application on MultiSenGE dataset

5

Contents

LULC and UF are mainly mapped using mono-temporal imagery and classical pixel approaches. This chapter answers the second sub-objective which assess of the synergy of multitemporal and multimodal satellite imagery provided by the Sentinel-1&2 satellite constellations for UF and LULC mapping. As a reminder, our hypothesis is that even at 10-meter spatial resolution, the intra-annual dynamics of vegetation should enable us to better discriminate LULC and in particular UF defined by a certain spatial organization but also portion of vegetation. Semantic segmentation preserves the spatial aspect of patches and state-of-the-art methods allow the use of multi-temporal data. With these approaches, the spatial organization of UF is then preserved. A spatio-temporal and spatio-spectral feature fusion method has been developed and applied on the MultiSenGE for UF and LULC classes dataset. The work was carried out on the whole Grand-Est region after the complete release of the OCSGE2 reference data at the beginning of 2022.

This chapter has been published in **Remote Sensing - MDPI**, in a Special Issue «*Advances in Deep Learning Techniques for the Analysis of Remote Sensing Time Series*» [Wenger et al., 2023b].

To complete this research work, an aditionnal test on the interest to add more dates for the S1 time series is presented in the last section of this chapter.

5.1 Introduction

Continental mapping of land cover is important in the context of global climate change. Complex workflows, often based on mono-temporal aerial or satellite imagery, can take many years to produce global land cover map. High resolution and frequently updated land cover maps became relevant to produce indicators to monitor and understand natural and anthropic processes. Moreover, at present, almost all LULC products (OSO [Inglada et al., 2017], ESA's World Cover 2020 [Zanaga et al., 2021], Google's Dynamic World [Brown et al., 2022] or Esri's 2020 Land Cover [Karra et al., 2021]) derived from automatic classifications based on classical machine learning method are describing urban areas only in one to four classes, with results that include salt and pepper effects [Guo et al., 2020]. However, urban areas also include vegetative areas, which are rich in biodiversity and provide urban cool islands [Yang et al., 2017]. Having an accurate and up-to-date land cover map where urban thematic classes are not reduced to two classes (urban/not urban) or four urban classes (road networks, dense and sparse built-up areas, and specialized areas) [El Mendili et al., 2020] is a major challenge, especially in the context of global change. Current works mapping urban areas in more than five classes often use very high resolution spatial imagery (e.g., Worldview-3), which is expensive with low temporal resolution (few images over time). This frequent update is relevant for urban planning and change detection analysis.

Many spatial programs offer images with a high temporal resolution, which allows obtaining relevant information on the temporal variability of anthropic and natural objects on the territory. This is particularly the case for the Copernicus program, developed by ESA (European Space Agency), with the Sentinel sensors, which allow the acquisition of the same site every six days, depending on the sensor. The data published per day for this spatial program corresponds to more than 15 TB of images from multiple sensors (Sentinel-1, Sentinel-2, Sentinel-3, and Sentinel-5). Several methods were developed to use satellite image time series (SITS) for land cover mapping [Inglada et al., 2017], change detection [Li et Narayanan, 2003], tree species detection [Karasiak et al., 2017], or crop classification [Li et al., 2015]. Thanks to the high revisiting period, Sentinel imagery (whether SAR for Sentinel-1 or optical for Sentinel-2), many works showed the importance of these images for Land Use/Land Cover (LULC) mapping both for optical [Praticò et al., 2021], and SAR imagery [Zhou et al., 2018]. Classical machine learning methods have reached their limit in terms of performance, and require a lot of exogenous indexes to achieve great results. The evolution of cloud computing has allowed the remote sensing community to develop new techniques to produce LULC maps based on classification methods and

more particularly on deep learning [Ma et al., 2019]. Many works used deep learning techniques and especially convolutional neural networks (CNN) for LULC mapping, either with pixel classification [Pelletier et al., 2019] or semantic segmentation [Zhang et al., 2018, Hafner et al., 2022]. Furthermore, there exist encoder/decoder networks such as U-Net [Ronneberger et al., 2015] or SegNet [Badrinarayanan et al., 2017] (also known as "U-Shape like" networks), which show excellent results for semantic segmentation or scene classification problems [El Mendili et al., 2020, Wenger et al., 2022c, Rußwurm et al., 2019].

Optical (Sentinel-2) and SAR (Sentinel-1) imagery come with complementary information on landscape elements. The first one describes the properties of surface materials and the second one provides the structural characteristics of landscape objects [Ienco et al., 2019]. The combination of optical and SAR imagery has led the community to develop methods to effectively perform their fusion. Three types of fusion classically exist: fusion at the pixel level (a), fusion at the feature level (b), and decisional fusion (c), which applies to the output of classification models [Ghassemian, 2016]. For classical machine learning approaches (i.e., Random Forest, SVM), data fusion consists of either concatenating the model input data (a) [Hedayati et Bargiel, 2018] or merging the probabilities (c) using decisional fusion algorithms [Inglada et al., 2015] such as majority voting of Dempster-Shafer [Smets et al., 1994]. These methods also allow the fusion of multimodal data, especially optical and SAR imagery [Clerici et al., 2017], and show a significant performance gain compared to the use of a single sensor. On the other hand, these methods treat these two modes of acquisition separately, and do not investigate the complementarity of the two acquisition modes and the multitemporal information.

Semantic segmentation and "U-Shape like" networks can be modified to perform feature fusion, which consists of combining features from several branches or layers of a network [Wenger et al., 2022c, Audebert et al., 2018]. The combination of optical and SAR imagery for feature fusion has been experimented with in several works, both for change detection [Liu et al., 2018] and for LULC classification [Kussul et al., 2017] by stacking two data sources (in this previous case, Sentinel-1 and Landsat-8 imagery). With the arrival of recurrent neural networks (RNN), it becomes possible, in addition to multimodal fusion, to take into account the multitemporal information of satellite images. These methods, called ConvLSTM, allow the use of both spatial and temporal dimensions for the extraction of spatio-temporal features. They have been widely used in multiple application fields, such as analyzing various video frame sequences [Pfeuffer et al., 2019], precipitation forecasting [Shi et al., 2015] or travel demand prediction [Wang et al., 2018]. In the field of remote sensing, ConvLSTM architectures have been proven to work well for Land Cover map-

ping [Ienco et al., 2019], change detection [Jamaluddin et al., 2021], deforestation mapping [Masolele et al., 2021], or rice field classification [Chang et al., 2022].

In a previous work [Wenger et al., 2022c], we experimented with feature fusion methods between Sentinel-2 single date optical imagery and spectral and textural indices for mapping urban areas in five thematic classes and obtained very promising results. To our knowledge, very few works attempt to classify urban areas with so many classes. Thus, the objective of this work is to explore the combination of multitemporal and multimodal imagery for urban fabric (UF) mapping using semantic segmentation networks. We make the hypothesis that (1) multitemporal optical and SAR imagery and (2) balancing the dataset can improve UF classification.

The rest of the paper consists of four parts: the choice of study sites and preprocessing methods in Section 5.2, the deep learning architecture used in Section 5.3, the analysis of the results in Section 5.4. The results are discussed in Section 5.5 and a conclusion and perspectives are detailed in Section 5.6.

5.2 Materials and Preprocessing Methods

5.2.1 MultiSenGE dataset

MultiSenGE [Wenger et al., 2022a] is a multitemporal and multimodal dataset developed over the Grand-Est region (Figure 5.1) in France. It covers 14 Sentinel-2 tiles over one of the biggest regions in France (57,433 km^2).

The dataset contains 8157 multitemporal patches of 256×256 pixels for the Sentinel-1 and Sentinel-2 sensors for 2020. A reference data, preprocessed from a Land Use Land Cover database (BDOCGE2), is included with each Sentinel-1 and Sentinel-2 patch to form data triplet. The global process of MultiSenGE construction can be found in Wenger et al. [2022a]. This dataset is one of the first providing multitemporal and multimodal imagery using Sentinel sensors for LULC applications. Furthermore, the reference data typology offers a diversity, especially for UF in 5 classes (see semantic classes typology in Figure 5.1).
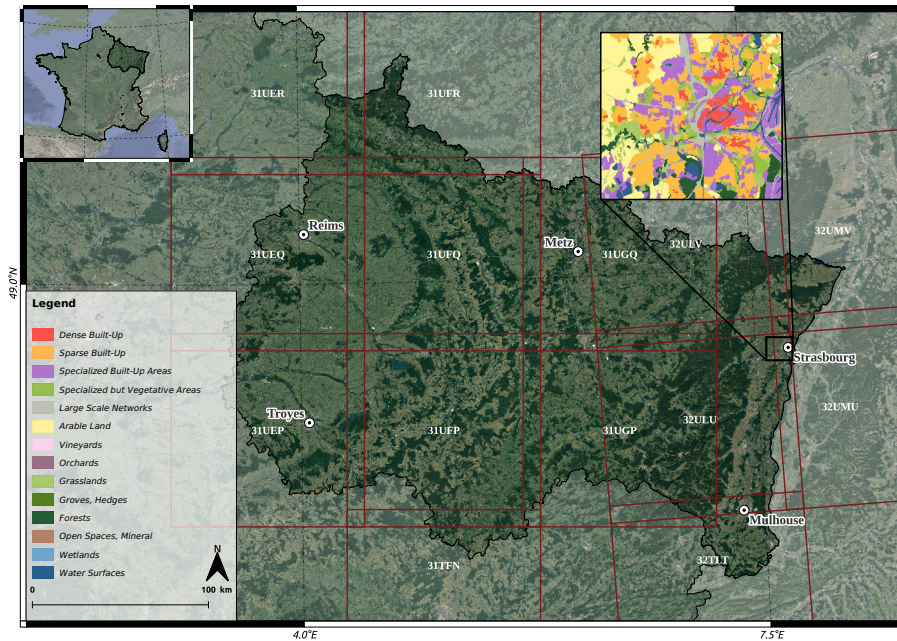
Figure 5.1 – *Grand-Est region with ground reference subset.*

Sentinel-1 carries a C-band SAR sensor and offers dual polarization data in Ground Range Detection (GRD) and Single Look Complex (SLC). Only GRD products are used in the construction of MultiSenGE and each Sentinel-1 patch consists of a stack of VV and VH bands.

Sentinel-2 images are acquired through the Theia land services and datacenter download portal (https://www.theia-land.fr/, accessed on 27 May 2022) and are L2A level, corrected for atmospheric effects, and accompanied by a cloud mask. Unlike Sentinel-1 products, where all available images of the series were downloaded, only images with less than 10% cloud cover are selected. Each MultiSenGE Sentinel-2 patch is composed of a stack of 10-m spectral bands (B2, B3, B4, B8) and 20-meter spectral bands (B5, B6, B7, B8A, B11, and B12) resampled to 10 m spatial resolution.

5.2.2 Optical and SAR Multitemporal Patches Selection

MultiSenGE [Wenger et al., 2022a] provides a set of functions to extract multiple Sentinel-1 and Sentinel-2 time series information for each patch. For example, it is possible to extract all patches that have at least one Sentinel-2 image associated for several months. Thus, we chose to explore the dataset to find the best compromise between temporal and spatial diversity. From Figure 5.2, we note that the dataset contains few images for winter and early spring.
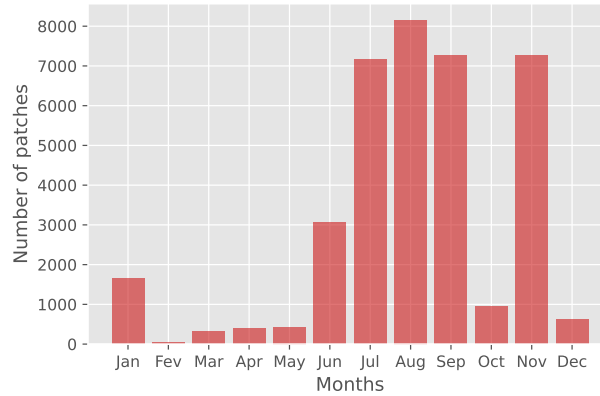
Figure 5.2 – *Number of patches with at least one Sentinel-2 image associated per month.*

Therefore, we decided to select patches for July (07), August (08), September (09), and November (11) to have the highest number of patches (Figure 5.2). To obtain images with a regular time-lapse, a constraint on the number of days between two consecutive months is applied to select the patches. To help in the choice of the best configuration of number and spatial diversity on a large region, we propose a web-page (http://romainwenger.fr/visu-multisenge/index.html, accessed on 5 September 2022) allowing displaying it by mapping the center of the patch (Figure 5.3).

We decided to select 17 days between two consecutive months to maximize the number of patches available (Table 5.1) as recommended in Jamaluddin et al. [2021]. In a previous work, some tests were made on the contribution of a larger number of Sentinel-2 dates. The first results showed that four dates without clouds are relevant to reduce uncertainties in a classification result compared to a larger number of dates which can increase them [Wenger et al., 2023a]. Moreover, the choice of this gap between two consecutive months is the best compromise between the number of patches and the spatial distribution of patches.
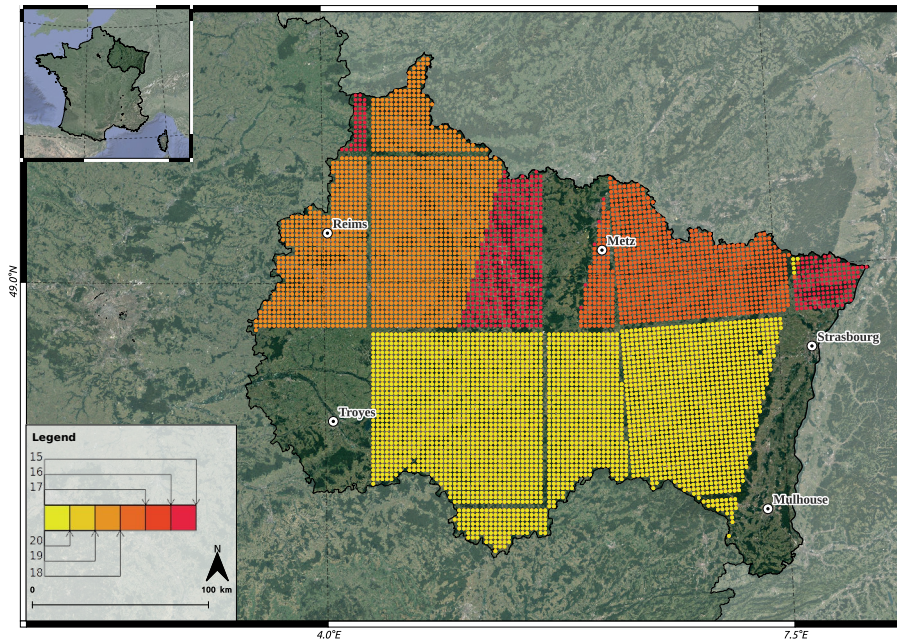
Figure 5.3 – *Distribution of patches over the study area according to the number of days between two consecutive dates for the months of July, August, September and November.*

Day gap for two consecutive months	Number of patches
15	6,560
16	5,890
17	5,890
18	4,960
19	3,178
20	3,178

Table 5.1 – *Number of patches depending on the day gap for two consecutive months.*

A sampling of the training, validation, and test sets is done according to a geographical stratification [Maxwell et al., 2021b] following Sentinel-2 tiling (Figure 5.4). Patches from tiles T31UFP and T31UGP are chosen for the validation set, T31UEQ for the test set, and all other available tiles for the training set (T32UMV, T32ULU, T32TLT, T31UGQ, T31TFN, T31UFQ, T31UFR). In total, there are 3369 patches for training (before data augmentation), 1911 patches for validation, and 610 patches for the test set.

Particular attention is accorded to keep the proportion of classes in the training and validation datasets. The patches from the test set are centered on Reims, another large city in the west of the region, allowing us to assess the performance of the model for urban thematic classes.
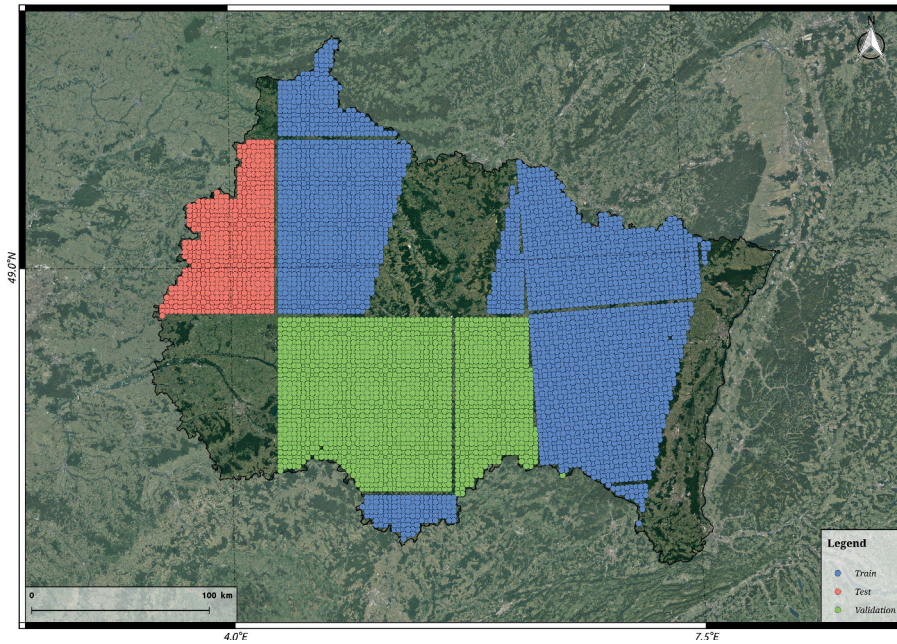
Figure 5.4 – *Train, validation and test sets for the selected multitemporal and multimodal patches.*

5.2.3 Reference Data Typology

The reference land cover dataset of MultiSenGE is described in 14 semantic classes. Following the choice of the different sets by geographical stratification, some classes are not homogeneously distributed in all the sets; for instance Orchards (8), Groves and Hedges (10), Open Spaces, Mineral (12), and Wetlands (13) represent mostly under 1% for the total dataset surface (Table 5.2).

MultiSenGE semantic classes	MultiSenGE distribution
Dense Built-Up (1)	0.37%
Sparse Built-Up (2)	3.64%
Specialized Built-Up Areas (3)	2.17%
Specialized but Vegetative Areas (4)	0.44%
Large Scale Networks (5)	0.91%
Arable Lands (6)	38.73%
Vineyards (7)	0.98%
Orchards (8)	0.15%
Grasslands (9)	18.87%
Groves, Hedges (10)	0.01%
Forests (11)	32.52%
Open Spaces, Mineral (12)	0.01%
Wetlands (13)	0.31%
Water Surfaces (14)	0.89%

Table 5.2 – *Semantic classes distribution for MultiSenGE dataset*

To reduce the unbalanced distribution of classes, we decided to merge some of them: (7) and (8) into a Vineyards and Orchards class, (10), (11) and (12) to create a class with Forests and semi-natural areas, and (13) and (14) for all the Water Surfaces (Table 5.2). Two different groupings are proposed on the baseline data, the first with 10 classes and the second with 6 classes to increase the number of thematic classes into several UFs (Table 5.2.3). The assumption is that with 10 land cover classes, the urban surfaces will be much better classified than with 6 land cover classes because the confusion between the natural classes will be reduced.

Even with these typologies in 6 or 10 classes, there are still some unbalanced classes (with less than 1% of the total land cover), especially for UFs classes (Dense Built-Up(1), Specialized but Vegetative Areas (4) and Large Scale Networks (5)). We have chosen not to merge these urban classes as they have already been used in several existing works [Inglada et al., 2017, El Mendili et al., 2020, Wenger et al., 2022c;a]. Indeed, these UFs semantic classes are often useful for urban planning, and decision-makers and are generic enough to map western cities.

Table 5.3 – *Semantic classes for MultiSenGE dataset and our reclassification in 6 and 10 classes.*

MultiSenGE semantic classes	10 classes	6 classes
Dense Built-Up (1)	Dense Built-Up (1)	Dense Built-Up (1)
Sparse Built-Up (2)	Sparse Built-Up (2)	Sparse Built-Up (2)
Specialized Built-Up Areas (3)	Specialized Built-Up Areas (3)	Specialized Built-Up Areas (3)
Specialized but Vegetative Areas (4)	Specialized but Vegetative Areas (4)	Specialized but Vegetative Areas (4)
Large Scale Networks (5)	Large Scale Networks (5)	Large Scale Networks (5)
Arable Lands (6)	Arable Lands (6)	Non urban areas (6)
Vineyards (7)	Vineyards and Orchards (7)	
Orchards (8)		
Grasslands (9)	Grasslands (8)	
Groves, Hedges (10)	Forests and semi-natural areas (9)	
Forests (11)		
Open Spaces, Mineral (12)		
Wetlands (13)	Water Surfaces (10)	
Water Surfaces (14)		

5.3 MODELS

In this section, we explain the two architectures (Figure 5.5) used for the different experiments. The first method uses a ConvLSTM module allowing the extraction of spatio-temporal features and takes as input the Sentinel-1 and Sentinel-2 multitemporal series (Section 5.3.1). The second method is an extension of the first one with the addition of a naive inception module for the extraction of spatial features based on filters of three different sizes in Section 5.3.2. Features computed for the two models are then concatenated and added as an input to a U-Net to obtain a LULC classification. These two methods were compared by taking as input different parameters described below (Section 5.3.3). Furthermore, evaluation metrics used in this study are presented at the end of the section (Section 5.3.5).
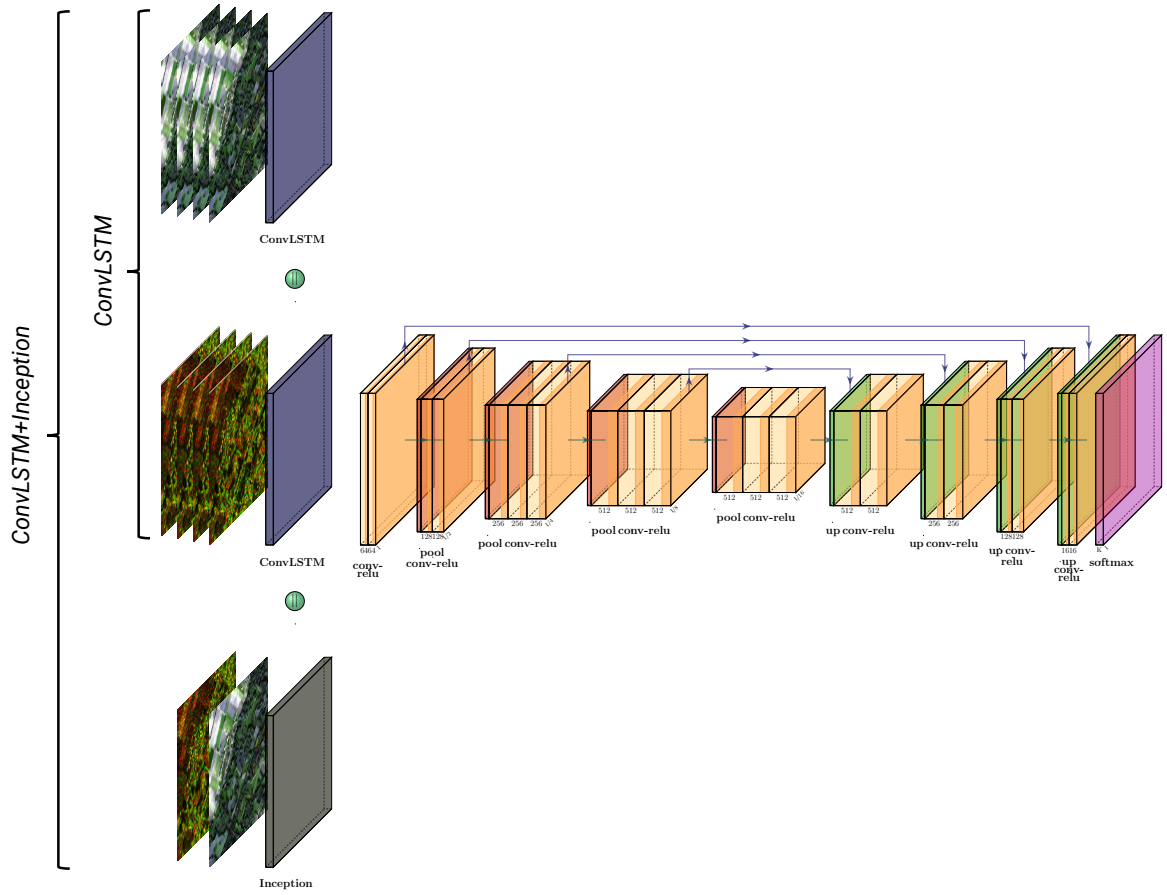
Figure 5.5 – *Sentinel-1 and Sentinel-2 ConvLSTM+Inception method (|| sign means concatenate). Inception module has been added and the U-Net network take as input the concatenation of the 2 ConvLSTM and the Inception module. This network is used for ConvLSTM and ConvLSTM+Inception methods.*

5.3.1   Spatio-Temporal Feature Extractor: ConvLSTM-S1/S2

The first method, *ConvLSTM* (Figure 5.5), is implemented by taking as the primary layer a ConvLSTM to extract spatio-temporal features that will be taken as input to a U-Net network [Ronneberger et al., 2015, Yakubovskiy, 2019]. The ConvLSTM layer is an extension of the LSTM which only computed temporal features without taking into account the spatial information of the 2D data. It is then that the ConvLSTM layer was set up which takes in input 5D data of the following form:

$$X_n \times T \times R \times C \times C' \tag{5.1}$$

where X_n represents the nth image, T the temporal dimension, R the number of rows, C the number of columns and C' the number of channels.

We used 256×256 patches with a temporal depth of 4 and 10 spectral bands. Our input data will therefore be of the following form:

$$X_n \times 4 \times 256 \times 256 \times 10 \tag{5.2}$$

The general structure of ConvLSTM (Figure 5.6) consists of taking as input $X_1, \ldots, X_t$ and returning as output spatio-temporal features which are 4D tensors. In Equation (5.3), which describes the ConvLSTM layer, $*$ denotes the convolutional operator and $\odot$ the Hadamard product [Shi et al., 2015].

$$
\begin{aligned}
i_t &= \sigma \left(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \odot C_{t-1} + b_i \right) \\
f_t &= \sigma \left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \odot C_{t-1} + b_f \right) \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tanh \left(W_{xc} * X_t + W_{hc} * \mathcal{H}_{t-1} + b_c \right) \\
o_t &= \sigma \left(W_{xo} * X_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \odot C_t + b_o \right) \\
\mathcal{H}_t &= o_t \odot \tanh \left(C_t \right)
\end{aligned}
\tag{5.3}
$$

In our case, we used a kernel of 3×3 and a filter size of 32. This layer is used for the Sentinel-1 time series, the Sentinel-2 time series, and finally for the two SAR and optical series together with a concatenation of the spatio-temporal features from each of the branches before input to the U-Net [Ronneberger et al., 2015] model.
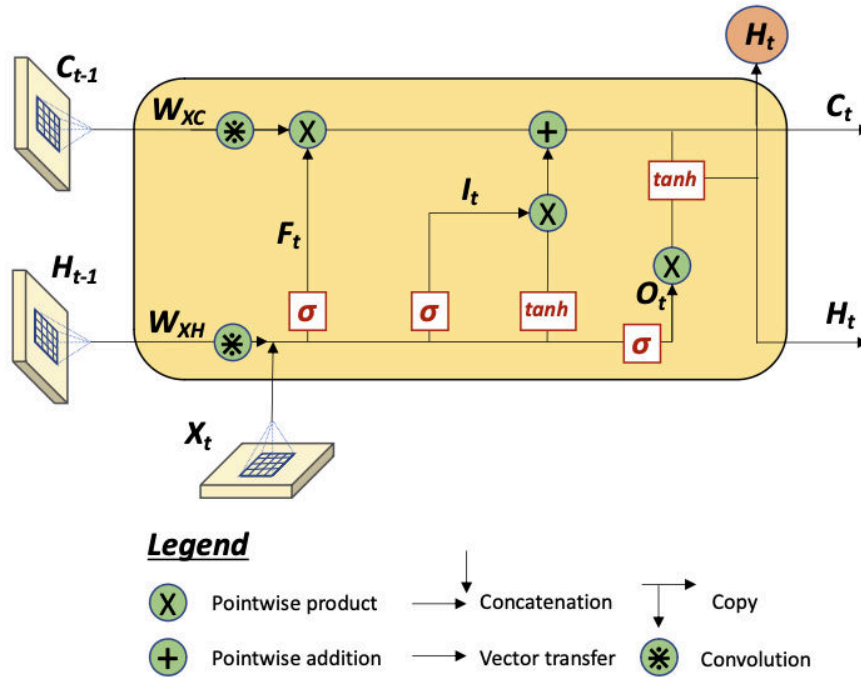
Figure 5.6 – *ConvLSTM structure.*

We chose to use a U-Net which is widely used for semantic segmentation with or without changes in the basic architecture [Zhang et al., 2021, Wei et al., 2021, Neves et al., 2020]. This network has the particularity to reduce the spatial information for the contracting part while increasing the features and combining the spatial and geographical information for the expensive part. The first part consists of a succession of convolutions followed by a ReLu (Rectified Linear Unit, $f(x) = max(0, x)$) and a MaxPooling operation. The second

part of the network is also composed of a series of convolutions, but this time it is followed by an UpSampling layer. At each pass through the network, the spatial resolution is initially reduced thanks to the downsampling layers, while the "spectral" information is increased. A second time, the "spectral" information is reduced to gain spatial information thanks to the UpSamplaing layer. VGG-16 has been chosen as a backbone because it is a good compromise between the complexity and the size of the network to limit overfitting. At the end of the network, a *softmax* function calculates the probability of each pixel. This network has been implemented thanks to Yakubovskiy [2019].

5.3.2 Spatio-Spectral-Temporal Feature Extractor: ConvLSTM+Inception-S1S2

The second method, *ConvLSTM+Inception* (Figure 5.5), consists of adding, in addition to the ConvLSTM modules, a Naive Inception module (Figure 5.7) which allows performing three 2D convolutions with filters of 1×1, 3×3 and 5×5 followed by a MaxPooling which allow limiting the overfitting and to save inputs in the model. Each 2D convolution of the model is followed by a ReLu (Rectified Linear Unit, $f(x) = max(0, x)$) at the end of the 2D convolution operations, the features extracted from the module are concatenated as well as the spatio-temporal features extracted from the ConvLSTM Sentinel-1 and Sentinel-2 modules. This second method applies only to the Sentinel-1 and Sentinel-2 time series for each ConvLSTM module as well as the first date of the series, in this case, the first date of July for the optical and SAR modules, which are concatenated before being passed into the Naive Inception module to extract spatio-spectral features.
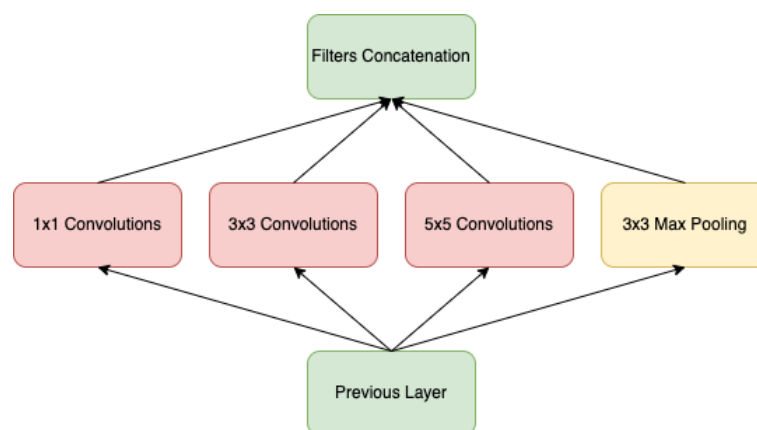
Figure 5.7 – *Naive Inception module.*

The U-Net used in the first part was also used in this one by taking as input the stack of features computed by the Inception module and the two ConvLSTM layers. This architecture was used for two experiments with, respectively, 6 and 10 classes as described in Section 5.2.3.

Table 5.4 – *List of experiments based on the methods presented.*

Name	Sensors	Method	Number of classes
ConvLSTM-S1	Sentinel-1	ConvLSTM	6 classes
ConvLSTM-S2	Sentinel-2	ConvLSTM	6 classes
ConvLSTM-S1S2	Sentinel-1 and Sentinel-2	ConvLSTM	6 classes
ConvLSTM-S1	Sentinel-1	ConvLSTM	10 classes
ConvLSTM-S2	Sentinel-2	ConvLSTM	10 classes
ConvLSTM-S1S2	Sentinel-1 and Sentinel-2	ConvLSTM	10 classes
ConvLSTM+Inception-S1S2	Sentinel-1 and Sentinel-2	ConvLSTM and Inception	6 classes
ConvLSTM+Inception-S1S2	Sentinel-1 and Sentinel-2	ConvLSTM and Inception	10 classes

5.3.3 Experimentation Details

Four main experiments (Table 5.3.3) are developed to test the contribution of each sensor for spatio-temporal feature extraction and an additional one that adds the spatio-spectral feature extractor combined with the spatio-temporal extractor. These tests are run for both 6 classes and 10 classes to explore the influence of a more diverse dataset which would offer less confusion and a better classification of both UFs and natural classes.

5.3.4 Implementation Details

Due to the imbalance of the dataset (Table 5.2), we chose to implement a Weighted Categorical Cross Entropy allowing us to assign a higher weight to the less balanced classes. The weight of each class is defined as the inverse of the frequency of the class [Wenger et al., 2022c, Audebert et al., 2018] and is commonly used in multiclass remote sensing classification [Ienco et al., 2017, Zhu et al., 2017b]. As demonstrated by Bai et al. [2022], Categorical Cross Entropy loss performs better than some other loss functions for semantic segmentation tasks. This method allows forcing the network to pay more attention to the less represented classes in the reference data (e.g., Built-Up (Class 1), Specialized But Vegetative Areas (Class 4), or Large Scale Networks (5)). Furthermore, Adam is selected as an optimizer [Kingma et Ba, 2014] as it performs better in remote sensing data than all others [Bera et Shrivastava, 2020, Pires de Lima et Marfurt, 2019, Abdollahi et al., 2021, Li et al., 2022].

The Sentinel-1 and Sentinel-2 data are normalized to the multi-temporal information

of each band using the following formula:

$$n = \frac{b - \bar{b}}{\sigma_b} \tag{5.4}$$

where n represents the normalized spectral band, b the reflectance values of each multi-temporal spectral bands, $\bar{b}$ the mean of the multitemporal reflectance values, and σ_b the standard deviation of the multitemporal reflectance values.

We chose to implement three different methods of data augmentation based on either 90, 180, or 270-degree rotation up/down flips and left/right flips. The training dataset is augmented to 75%. EarlyStopping, present in the Keras (https://www.tensorflow.org/api_docs/python/tf/keras accessed on 27 May 2022) library, is used with a patience of 20 epochs to avoid any overfitting. Adam optimizer is used with a LR of 10^{-3} and we reduce it by a factor of 0.1 after 5 epochs each time a plateau is reached thanks to the ReduceLROnPlateau method. Every Python code were run on a GPU cluster using 3 RTX6000 with 24 GB of VRAM each (72 GB in total). This allowed us to use a batch size of 16 for each run proposed in the paper.

5.3.5 Evaluation Metrics

Three evaluation metrics are used to assess the overall performance of the models and each class studied: $F1_{Score}$, Recall, Precision [Maxwell et al., 2021a] and Cohen's Kappa.

Precision score, also known as User's Accuracy, allows extracting the number of correctly classified pixels in the classified image and is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5.5}$$

Recall score, also known as Producer's Accuracy, allows extracting the percentage of well-predicted positives compared to all positives and is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{5.6}$$

$F1_{Score}$, also known as Dice, is the harmonic mean between the two previously explained metrics, Precision and Recall. It is calculated as follows :

$$F1_{Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5.7}$$

We also chose to compute each weighted metric for all classes to globally evaluate each model. They are calculated taking the mean of all class while considering each class's support.

Cohen's Kappa measure the level of agreement between two annotations.

$$\kappa = \frac{P_o - P_e}{1 - P_e}.$$

where P_o defined the empirical probability of agreement (also known as the observed agreement ratio) and P_e the expected agreement.

5.4 RESULTS

This section presents the results obtained for the two methods developed over the test dataset. The test data set is independent of the training set and the validation set and includes all available patches for the T31UEQ tile, as seen in Section 5.2.2. First, we present the results for six LULC classes in Section 5.4.1 then the results for 10 LULC classes in Section 5.4.2 and finally we compare UFs classification results between 6 and 10 LULC classification methods for UFs in Section 5.4.3.

5.4.1 6 Classes Results

All the results for the six semantic classes are compiled in Table 5.5. For these first experiments, we study the influence of the addition of different sensors (Sentinel-1 and Sentinel-2), for one date per month for July, August, September, and November. We notice that with 6sixLULC classes, it is difficult for the tested methods to have a convergence of the scores for all the classes. We notice that the ConvLSTM-S1 method offers the best $F1_{Score}$ with 0.1344 for the classification of the *Specialized but Vegetative Areas (4)*.

	ConvLSTM-S1			ConvLSTM-S2		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.1335	**0.9397**	0.2337	0.2579	0.8704	0.3980
Class 2	0.4476	0.3809	0.4116	0.5575	0.7268	0.6310
Class 3	0.3560	0.5813	0.4416	0.3100	**0.7763**	0.4431
Class 4	**0.0775**	0.5072	**0.1344**	0.0528	0.4858	0.0953
Class 5	0.1313	0.5516	0.2122	0.2137	0.7995	0.3372
Class 6	0.9937	**0.8937**	**0.9410**	**0.9979**	0.8663	0.9274
W-Avg	0.9469	**0.8661**	0.9001	0.9544	0.8574	0.8958
	ConvLSTM-S1S2			ConvLSTM+Incep-S1S2		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	**0.3122**	0.7624	**0.4430**	0.2308	0.8599	0.3639
Class 2	0.5671	**0.7706**	**0.6533**	**0.6260**	0.6472	0.6364
Class 3	0.4654	0.6859	0.5545	**0.4794**	0.7647	**0.5894**
Class 4	0.0314	**0.5739**	0.0595	0.0312	0.4461	0.0584
Class 5	**0.2745**	**0.8085**	**0.4099**	0.2736	0.7898	0.4064
Class 6	0.9971	0.8446	0.9145	0.9965	0.8719	0.9301
W-Avg	0.9578	0.8369	0.8875	**0.9591**	0.8596	**0.9018**

Table 5.5 – *Results of all methods for the test zone located in the west of the Grand-Est region.*

The addition of multitemporal and multimodal data for the extraction of spatio-temporal features (ConvLSTM-S1S2 method) allows obtaining a large part of the best Recall score and $F1_{Score}$, especially for *Dense Built-Up (1)*, *Sparse Built-Up (2)* and *Large Scale Networks (5)*. The addition of the Inception module for the extraction of spatio-spectral features (ConvLSTM+Inception-S1S2) does not allow a significant improvement of the classes, even if in terms of Weighted-$F1_{Score}$, it is very close to the ConvLSTM-S1S2 method with 0.8875 against 0.9018 for the latter (Table 5.5 and Figure A.5). On the other hand, ConvLSTM+Inception-S1S2 obtains the best Weighted-Precision and Weighted-$F1_{Score}$ with 0.9591 and 0.9018. The confusion matrix (Figure A.6) informs us about a strong confusion between *Specialized But Vegetative Areas (4)* and *Non-urban areas (6)*, probably due to the imbalanced dataset because *Non-urban areas (6)* covers 92.46% of the dataset. Furthermore, it is complex to differentiate it from other natural classes as *Non-urban areas (6)* is the aggregation of all other natural areas. Moreover, we notice a strong confusion between *Dense Built-Up (1)* and *Sparse Built-Up (2)* which are complex classes to differentiate because their texture and spectral signature are very close at 10m. The only difference between these two

Table 5.6 – *Cohen's Kappa for each method for 6 semantic classes. (In bold the best method.*

Method	Cohen's Kappa
ConvLSTM-S1	0.3929
ConvLSTM-S2	**0.4223**
ConvLSTM-S1S2	0.3852
ConvLSTM+Inception-S1S2	0.4186

UF classes is the portion of vegetation between buildings Wenger et al. [2022c], El Mendili et al. [2020], which are close to none for *Dense Built-Up (1)* and are restricted to the personal garden for *Sparse Built-Up (2)*. We can see that Recall values are systematically higher than the Precision Values whatever the method. However, the Precision value is also always higher (classes *Sparse Built-Up (2)* and *Specialized Built-Up Areas (3)*) or very similar (classes *Dense Built-Up (1)*, *Specialized But Vegetative Areas (4)* and *Large Scale Networks (5)*) with ConvLSTM+Inception-S1S2 than other methods. As seen in Table 5.4.1 with the Cohen's Kappa metric, we can see that there is a very low agreement as seen with the scores between 0.3929 and 0.4223.

The visual analysis (Figure 5.8) is performed on three patches with different urban densities: 31UEQ_GR_7453_4112, 31UEQ_GR_6939_6682, 31UEQ_GR_3855_8481 (these patches can be viewed by downloading MultiSenGE Wenger et al. [2022a]) . These results confirm the statistical results where the best classifications are found for the ConvLSTM-S1S2 and ConvLSTM+Inception-S1S2 methods. We notice a better delimitation of the boundaries between UFs (class (1) to (5)) and *Non-urban areas (6)*.
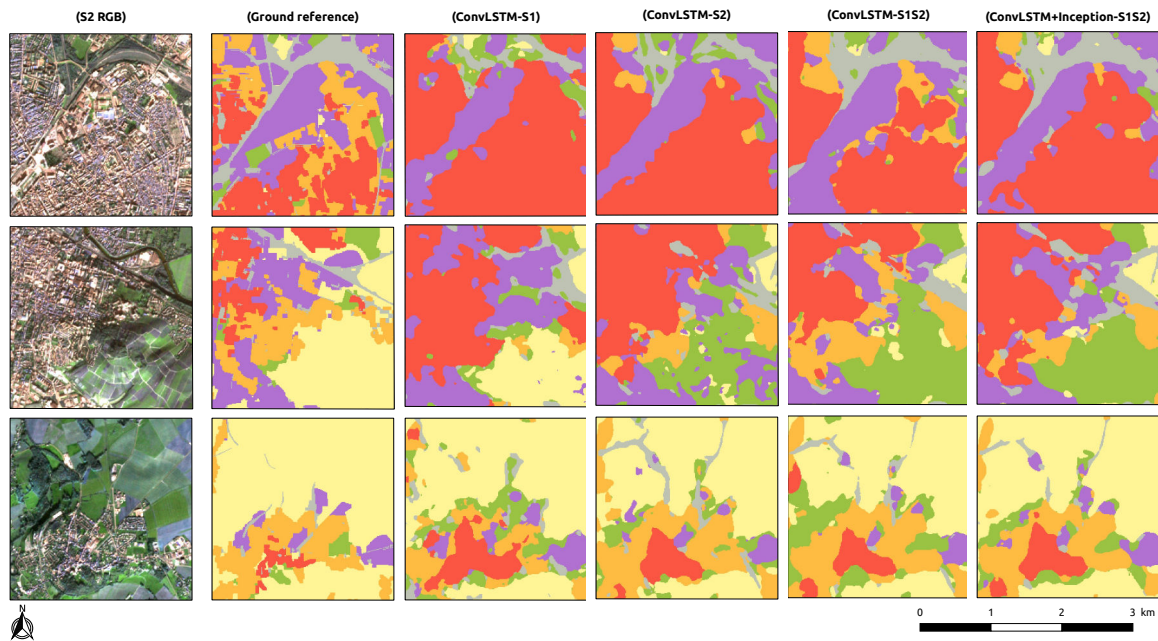
Figure 5.8 – *Results for each method for 6 semantic classes (Legend is available in Table 5.2.3)*

5.4.2 10 Classes Results

This section summarizes all quantitative and qualitative results for the 10 LULC classifications. Table 5.7 contains all the statistical results for the 10 classes of experiments. We notice that all the global evaluation metrics such as the accuracy, the Weighted-$F1_{Score}$ and the Mean-$F1_{Score}$ have the highest scores for the ConvLSTM+Inception-S1S2 method with 0.8831, 0.6373 and 0.8851, respectively. Moreover, the vast majority of the classes have a higher $F1_{Score}$ for the latter. Only three classes have higher scores for another method (ConvLSTM-S2): *Dense Built-Up (1)*, *Sparse Built-Up (2)* and *Vineyards and Orchards (7)*. Moreover, this method allows better extraction of natural classes probably thanks to the spatio-spectral feature extractor (Figure A.7). The analysis of the confusion matrixes (Figure A.8) allows us to identify weaker confusions for the ConvLSTM+Inception-S1S2 method than for all the other methods tested. We can also notice that Recall values are always higher than Precision for almost every UF classes and method. This trend does not apply to natural classes.

	ConvLSTM-S1			ConvLSTM-S2		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.1872	**0.9247**	0.3114	**0.5629**	0.4968	**0.5278**
Class 2	0.5718	0.5224	0.5460	**0.6814**	0.7625	0.7197
Class 3	0.4208	0.6480	0.5103	0.4909	0.7329	**0.5880**
Class 4	0.0892	**0.4973**	0.1512	0.1597	0.3499	0.2193
Class 5	0.2142	0.6183	0.3182	0.2914	0.8076	0.4283
Class 6	0.9649	0.8540	0.9060	0.9838	0.9263	0.9542
Class 7	0.8361	0.5625	0.6726	**0.9003**	0.8737	**0.8868**
Class 8	0.3890	0.4111	0.3997	0.5720	0.4336	0.4933
Class 9	0.7515	0.8280	0.7879	**0.9002**	0.7697	0.8299
Class 10	0.3143	0.4748	0.3782	0.1611	**0.9106**	0.2737
W-Avg	0.8422	0.7836	0.8055	0.9000	0.8517	0.8696
	ConvLSTM-S1S2			ConvLSTM+Incep-S1S2		
	Precision	Recall	F1	Precision	Recall	F1
Class 1	0.2736	0.8199	0.4103	0.3870	0.7190	0.5031
Class 2	0.6498	0.7287	0.6870	0.6672	**0.8066**	**0.7303**
Class 3	**0.5840**	0.3955	0.4716	0.4612	**0.7632**	0.5749
Class 4	**0.1885**	0.2692	0.2217	0.1863	0.3643	**0.2465**
Class 5	0.2739	**0.8666**	0.4163	**0.4290**	0.7560	**0.5474**
Class 6	**0.9862**	0.9033	0.9430	0.9718	**0.9558**	**0.9637**
Class 7	0.7822	**0.9203**	0.8457	0.8869	0.8512	0.8687
Class 8	0.4914	**0.4555**	0.4728	**0.7422**	0.3949	**0.5155**
Class 9	0.8516	0.8533	0.8524	0.8585	**0.8643**	**0.8614**
Class 10	0.2759	0.8660	0.4185	**0.4654**	0.7074	**0.5614**
W-Avg	0.8825	0.8482	0.8600	0.8977	**0.8831**	**0.8851**

Table 5.7 – *Results of all methods for 10 LULC classes the test zone located in the west of the Grand-Est region.*

As seen in Table 5.4.2, the best agreement between reference data and classification are for the ConvLSTM+Inception-S1S2 with 0.7945 for Cohen's Kappa evaluation metric. This confirmed the best results for this method compared to others.

To perform the qualitative assessment for this section, we used the same

Table 5.8 – *Cohen's Kappa for each method for 10 semantic classes. (In bold the best method.*

Method	Cohen's Kappa
ConvLSTM-S1	0.6422
ConvLSTM-S2	0.7445
ConvLSTM-S1S2	0.7482
ConvLSTM+Inception-S1S2	**0.7945**

patches as in Section 5.4.1 to compare the two approaches (31UEQ_GR_7453_4112, 31UEQ_GR_6939_6682, 31UEQ_GR_3855_8481).

The qualitative analysis allows us to observe that the *Vineyard and Orchards (7)* class initially strongly confused with the Specialized but vegetative Areas class (4) in the 6-class methods is correctly classified. The urban boundaries are correctly defined. We also notice that for the natural areas and more particularly the class *Forest and semi-natural areas (9)*, the boundaries between the classes are more precise for the ConvLSTM+Inception-S1S2 method than for all the other methods. For the second best performing method (ConvLSTM-S2), a strong confusion is found between this class and the *Water Surfaces (10)* class. It is also interesting to note that the small roads, initially not present in the reference data, are detected for the ConvLSTM-S2 and ConvLSTM+Inception-S1S2 methods (Figure 5.9).
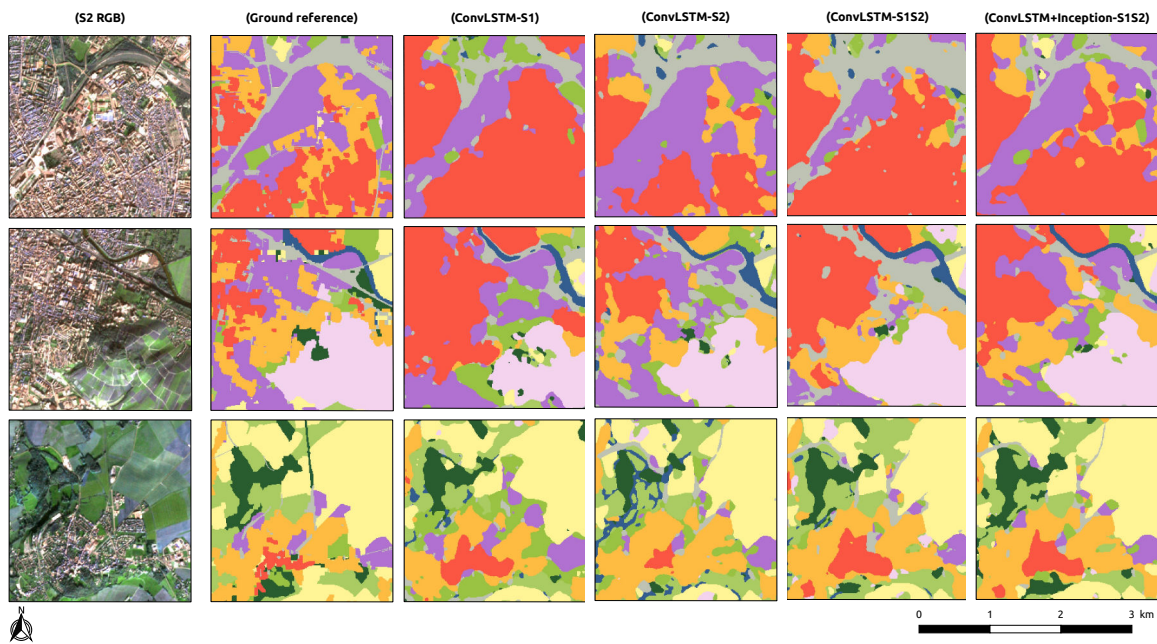
Figure 5.9 – *Results for each method for 6 semantic classes (Legend is available in Table 5.2.3).*

5.4.3 UFs Analysis

The evaluation metrics $F1_{Score}$ for each class UFs are displayed on several graphs (Figure 5.10) for the best 10-class method and all 6-class methods. We notice that the ConvLSTM+Inception-S1S2 method for 10 class LULC classes provides better results for all five class UFs chosen for this study. Only the *Specialized Built-Up Areas class (3)* performs better for ConvLSTM+Inception-S1S2 at 6 classes than at 10 classes. On the other hand, all other classes are better classified for 10 classes because the confusion between them is strongly reduced. As seen in Figures 5.8 and 5.9, the better performance with 10 LULC classes is particularly noticeable at the level of the urban periphery. ConvLSTM+Inception-S1S2 allows better separating the Dense *Built-Up (1)* and the *Sparse Built-Up (2)*. *Large Scale Networks (5)* were better extracted using ConvLSTM+Inception-S1S2 and less confusion can be seen with *Specialized Built-Up Areas (3)* compared to other methods tested.
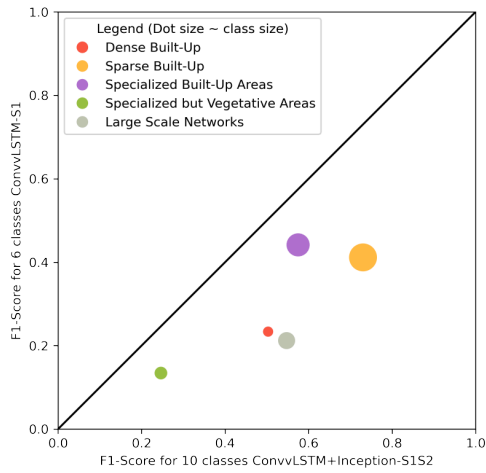
(a) 10 classes ConvLSTM+Inception-S1S2 vs 6 classes ConvLSTM-S1
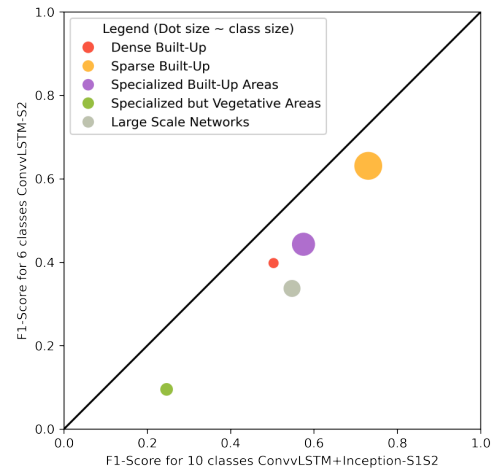
(b) 10 classes ConvLSTM+Inception-S1S2 vs 6 classes ConvLSTM-S2
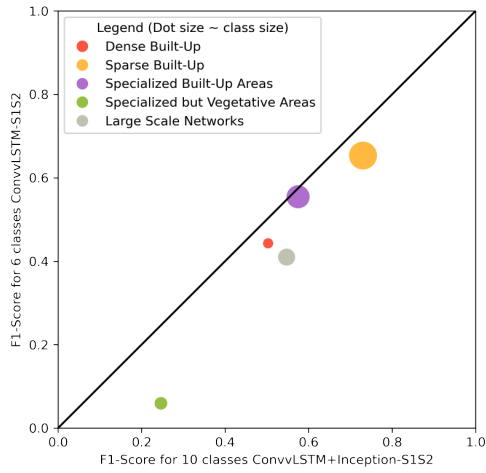
(c) 10 classes ConvLSTM+Inception-S1S2 vs 6 classes ConvLSTM-S1S2
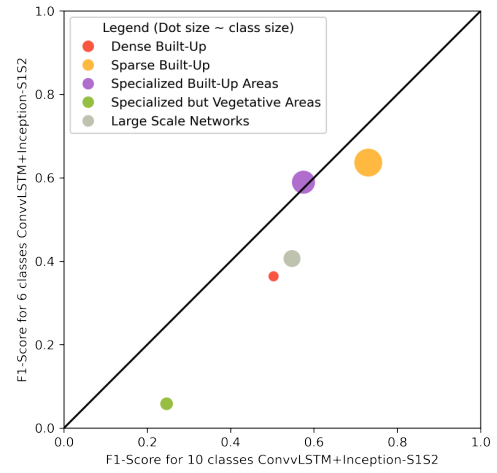
(d) 10 classes ConvLSTM+Inception-S1S2 vs 6 classes ConvLSTM+Inception-S1S2

Figure 5.10 – *Scatter plot to compare UFs (classes 1 to 5 in Table 5.2.3) classifications of ConvLSTM+Inception-S1S2 for 10 classes between every methods implemented for 6 classes*

5.5 DISCUSSION

In this paper, we developed a semantic segmentation method taking as input multitemporal and multimodal imagery from the MultiSenGE dataset. We propose a network consisting of two extractors, one for spatio-spectral features and the other for spatio-temporal features. The latter provides the best results compared to the other tested approaches. Moreover, discriminating an over-represented class improves the classification results for the studied object, in our case the UFs, reducing both intra and inter-classes confusion.

5.5.1 Application on UF Mapping

For UF mapping, results showed higher quantitative metrics at 10 classes using SAR and optical time series and an Inception module allowing the extraction of spatio-spectral features in addition to spatio-temporal features. This latter approach allows reducing the strong confusion between classes coming from imbalanced datasets during the 6 LULC classification. This was particularly true for Figures 5.8 and 5.9 where *Vineyards and Orchards (7)* were strongly confused with *Specialized but Vegetative Areas (4)*, a class that also includes scattered trees in urban parks or squares. Moreover, the results presented allowed us to see the contribution of Sentinel-1 SAR imagery thanks to the better detection of natural surfaces in the periphery of the UF. Without this acquisition mode, the confusion between the natural classes and *Specialized but Vegetative Areas (4)* would probably have been more important and would not have allowed a better detection of these areas.

Sentinel-2 satellite imagery provides a high spatial and high temporal coverage thanks to its temporal resolution (3 to 6 days). However, UF mapping remains challenging as these classes contains very small objects with various spectral diversity. Indeed, the distinction between UFs classes is mainly based on the amount of vegetation in each class (e.g., *Dense Built-Up (1)* and *Sparse Built-Up (2)*). On the other hand, temporal diversity provides essential information to refine the classifications as it offers the possibility to assess the evolution of the landscape and especially the vegetation through time. Results obtained in this study are superior to existing work mapping UF in several classes from Sentinel imagery [El Mendili et al., 2020, Wenger et al., 2022c].

5.5.2 Comparison with a State of the Art LULC Product

Compared to existing products such as OSO [Inglada et al., 2017], our method is better to describe UF classes on most of the class. Indeed, using semantic segmentation instead of classical machine learning approaches (e.g., Random Forest) reduces the salt and pepper effect. Furthermore, we include a fifth class, *Specialized but Vegetative Areas (4)*, which is almost never mapped in existing works using 10 m spatial resolution imagery (e.g., Sentinel). In fact, OSO only derives UFs in four classes, *Dense Built-Up*, *Sparse Built-Up*, *Specialized Built-Up*, and *Large Scale Networks*. For natural areas, OSO has 13 semantic classes, which is slightly higher than our approach. However, we are only experimenting with a specific region in France and a regional LULC semantic segmentation dataset, which reduces the possibilities to extend the number of classes. Due to the complex spectral diversity of the *Specialized but Vegetative Areas (4)* class because of the large number of objects (Trees, Grasslands, Minerals . . . , also included in other LULC classes), $F1_{Score}$ cannot exceed 0.25.

However, almost every vegetative areas inside urban areas remains to *Specialized but Vegetative Areas (4)* using our method. Confusions are mostly seen outside urban areas.

5.5.3 Network Performance

Semantic segmentation networks, and more specifically the encoder-decoder like structures, allow us to obtain a map with the spatial extent of each class (assigning to each pixel a label and taking into account the spatial context of the image). Thus, we can note a lack of precision in the border of the classes, in particular those of the UF. Moreover, the complexity and density of UF make the distinction between classes difficult, especially at 10 m spatial resolution. The geographical area, being anisotropic in nature, differences in the distribution and frequency of the classes over the territory also make the classification more complex and challenging. This could be assessed by doing pixel-wise classification and balancing each class in the dataset. Through this technique, salt and pepper noise, which was almost erased with semantic segmentation, could happen again. Weighting the loss is one of the methods that has been successful in the community to assess this challenge [Wenger et al., 2022c, Audebert et al., 2018, Zhu et al., 2017b, Ienco et al., 2017]. In the case of our study, it seems to be working because the least represented classes (less than 2% of all classes) reach F1-Scores above 0.5. However, this strategy seems to have difficulties for the least separable and most confused classes, such as *Dense Built-Up (1)* (often confused with *Sparse Built-Up (2)*), *Specialized but Vegetative Areas (4)* or *Grasslands (8)*.

Through these experimentations, ConvLSTM+Inception-S1S2 for 10 LULC classes appears to be the best method to map UF using multitemporal and multimodal imagery. Detailing an over-represented class allowed the network to improve the results by reducing intra-class confusion. Moreover, the contribution of an Inception module and of spatio-spectral features could be one of the reasons for the improvement of the classification results. The spatial context of an image, in semantic segmentation problems, is an important aspect in classification results. The addition of this module, allowing the calculation of features according to several filter sizes, may have contributed to the results obtained. On the other hand, according to the classification results, the spatio-spectral feature extractor provides important information on the smallest objects of the territory thanks to multiple kernel filter sizes.

5.6 Conclusions

In this study, we demonstrated the contributions of multitemporal and multimodal imagery and the use of deep learning models allowing the extraction of spatio-spectral and spatio-temporal features for a better extraction and semantic segmentation of UF. Furthermore, the results, which demonstrate better $F1_{Score}$ for the 10 classes ConvLSTM+Inception-S1S2 method, showed that it is better to segment and diversify an over-represented class composed of spectrally and texturally distinct objects. This method has also greater metrics scores thanks to the addition of a spatio-spectral feature extractor. The current scores for UF are encouraging and show that combining and extracting different types of features and balancing the initial dataset provides better results by reducing confusion between the classes studied (Figures 5.9, 5.10 and A.7). The developed methods could be used in large-scale LULC classification to study their genericity under different scenarios at different spatial scales (e.g., over France and/or Europe). For example, for cities slightly different than western cities, transfer learning could be applied and compared to a network trained from scratch. To increase the accuracy of the classifications, one of the perspectives could be to add an image with a very high spatial resolution (e.g., Pléiades or Spot6/7) and to merge the features from the multi-temporal optical and SAR images and very high spatial resolution. Furthermore, we would like to explore spatial and temporal inference to cover large territories and produce a land cover map that may be included in climatic models to characterize Local Climate Zone (LCZ).

5.7 Additional test with multitemporal S2 and S1 times series

During the experimental step of this study concerning the inputs in the ConvLSTM+Inception-S1S2 method, a question was to assess whether the inclusion of S1 time series would improve the result. In order to do this, we select 8 Sentinel-1 dates (the first date of each month between April 2020 and November 2020) and we keep 4 Sentinel-2 dates (one date per month with 17 days difference for July, August, September and November). This test has been named ConvLSTM+Inception-TS-S1S2 (Table 5.9).

Results of this additional experiments not presented in the paper, showed similar weighted $F1_{Score}$ between 4 Sentinel-1 dates and 8 Sentinel-1 dates for ConvLSTM+Inception-S1S2 method (Table 15.9). Most of the classes have a better F1Score for 4 dates than 8 dates.

In conclusion, this additional test showed that include time series of Sentinel-1 (SAR) images does not improve the mapping of LULC. Moreover, time processing is reduced in

term of pre-processing and computing model by using multi-temporal and multi-modal Sentinel imagery.

	ConvLSTM+Inception-TS-S1S2			ConvLSTM+Inception-S1S2		
	Precision	Recall	F1	Precision	Recall	F1
Dense Built-Up (1)	0.3638	0.7391	0.4876	0.3870	0.7190	0.5031
Sparse Built-Up (2)	0.6485	0.7856	0.7105	0.6672	0.8066	0.7303
Specialized Built-Up Areas (3)	0.4265	0.8157	0.5602	0.4612	0.7632	0.5749
Specialized but Vegetative Areas (4)	0.1590	0.3848	0.2250	0.1863	0.3643	0.2465
Large Scale Networks (5)	0.3424	0.8219	0.4834	0.4290	0.7560	0.5474
Arable Lands (6)	0.9803	0.9422	0.9609	0.9718	0.9558	0.9637
Vineyards and Orchards (7)	0.9098	0.8293	0.8677	0.8869	0.8512	0.8687
Grasslands (8)	0.6749	0.4191	0.5171	0.7422	0.3949	0.5155
Forests and semi-natural areas (9)	0.8767	0.8324	0.8539	0.8585	0.8643	0.8614
Water Surfaces (10)	0.3553	0.7877	0.4897	0.4654	0.7074	0.5614
W-Avg	0.8999	0.8713	0.8797	0.8977	0.8831	0.8851

Table 5.9 – *Results for ConvLSTM+Inception-TS-S1S2 and ConvLSTM+Inception-S1S2.*

Spatial and temporal inference

6

To evaluate the generalization capabilities of the multitemporal and multimodal method (ConLSTM+Inception-S1S2) developed in Chapter 5, we employed both spatial and temporal inference. Spatial inference involves using the model on a geographically different area from the training area, while temporal inference involves using the model on the same training area but for different dates. By using temporal and spatial inference, we make the hypothesis that we can detect artificial surfaces at a lower cost and generalize the results to a larger scale to produce a LULC map. This study responds to the third sub-objective which aims to evaluate the spatial and temporal inference capacity of the classification models to generalize the proposed approach for mapping large-scale territories and analyzing their dynamics.

This chapter focusing on spatial inference is accepted for an oral presentation at Joint Urban Remote Sensing Event 2023 (JURSE2023) in Heraklion, Greece. A short conference paper has been accepted and will be published in IEEE JURSE 2023 Proceedings.

Tests on temporal inference have been added in this chapter as a new contribution allowing to monitor LULC dynamics.

6.1 INTRODUCTION

For many years, urban sprawl and global warming have been at the heart of the scientific community's concerns. Indeed, by 2050, no less than three out of four inhabitants will be living in cities [United Nations Department of Economic and Social Affairs Population Division, 2018b]. This urban sprawl causes more and more pressure on ecosystems and consumes many natural areas essential for the preservation of biodiversity. In many cities, urban green areas are located in the private and public domain and are essential to maintain urban cool islands [Yang et al., 2017] especially in a context of global change. Therefore, urban planners need frequent and up-to-date Land Use Land Cover mapping in order to detect and quantify changes at the district or city level.

International space agencies have set up Earth observation programs with the deployment of many satellites, whether in passive remote sensing (optical) but also active remote sensing (radar). This is the case of the ESA (European Space Agency) and the Copernicus program who launched in orbit Sentinel constelation which produce every day several TB of data ($\sim$ 15 TB each day).

In order to process this increasing amount of data, researchers have developed methods based on artificial intelligence and more particularly neural networks. These methods need a large amount of training data to be efficient [Anaby-Tavor et al., 2020].

Semantic segmentation is one of the methods in the field of artificial intelligence. It assigns a semantic class to each pixel of an image and produce a mask according to a probability of belonging to each class [Ma et al., 2019]. Also, one of the main advantage of this method is its ability to recognize a set of categories that forms a cluster of pixel of the same class. It reduces the salt and pepper noise resulting from classical pixel approches [Hirayama et al., 2019]. This method is used to produce a Land Use Land Cover map [Kussul et al., 2017]. One of the most widely used semantic segmentation networks is the U-Net [Ronneberger et al., 2015] which has already proven its efficiency, especially in urban fabric semantic segmentation works [Wenger et al., 2022c, El Mendili et al., 2020].

Feature fusion has been widely used by the community to perform multitemporal and/or multimodal semantic segmentation [Hu et al., 2017]. Temporal dynamics of land cover objects are retrieve from multitemporal imagery and properties and structural characteristics are provided by optical and radar [Clerici et al., 2017] imagery. The synergy of these two aspects shows interesting results for natural areas [Betbeder et al., 2014] and urban fabric mapping [Iannelli et Gamba, 2018].

In previous work [Wenger et al., 2023b], it has been shown that the contribution of multitemporal and multimodal imagery improves urban fabric semantic segmentation.

Thus, a convolution network has been developed and trained on the MultiSenGE [Wenger et al., 2022b] dataset which was built for the entire Grand-Est region in France. It takes as input multitemporal Sentinel-1 and Sentinel-2 imagery. To perform large-scale mapping, three types of methods are used in the literature: classical training (selection of training and test data over the entire study area, usually applying geographic stratification), domain adaptation (training on two source and target areas with or without reference data for the target area) [Hafner et al., 2022], and fine-tuning (model pre-trained on the source area and then re-trained on the target area, which allows the weight to converge faster) [Pires de Lima et Marfurt, 2019].

The objective of this work is to explore (1) spatial inference by classifying five cities over France located in the same climate zone and assess the limit of spatial inference with other cities in Europe and Africa, and (2) temporal inference by classifying Grand-Est region for 2018 and 2022 to detect abrupt changes in urban fabric. Through this approach, we would like to explore the genericity of a deep learning model, both for spatial and temporal aspect without having to re-train the model.

6.2 METHODS

6.2.1 Datasets

In this work, we have chosen to use MultiSenGE [Wenger et al., 2022a] which is a Land Use Land Cover dataset developed over the entire Grand-Est region in France. It contains 8,157 multitemporal and multimodal patches (256×256) cut from the Sentinel-1 and Sentinel-2 time series which represents 14 Sentinel-2 tiles. This dataset was developed for the year 2020, which corresponds to the year of production of the reference data, OCSGE2-GEOGRANDEST[1], used to match the satellite data. It has also been reprocessed to be coherent with the spatial resolution of the satellite images.

The Sentinel-2 images used by MultiSenGE have been uploaded through the Theia portal[2]. They are offered to users at L2A level which corresponds to a correction of the atmosphere effects, slope effects and the availability of a cloud mask. This dataset used only images containing less than 10% cloud cover and each Sentinel-2 patch is composed of a stack of bands at 10m (B2, B3, B4, B8) and 20m (B5, B6, B7, B8A, B11 and B12) spatial resolution.

Sentinel-1 data were downloaded using the *s1tiling*[3] processing chain, developed by

---

[1]https://www.geograndest.fr
[2]https://www.theia-land.fr/
[3]https://github.com/CNES/S1Tiling

CNES (Centre National des Etudes Spatiales). Only the Ground Range Detection (GRD) products have been kept and the Sentinel-1 patches are composed of a stack at 10m spatial resolution of the VV and VH radar bands.

In previous work (Chapter 5), the reference data, initially in 14 classes, was reclassified into 10 classes by merging the least represented classes (Table 6.1).

Table 6.1 – *Semantic classes for MultiSenGE and 10 classes classification (adapted from [Wenger et al., 2023b])*

MultiSenGE semantic classes	10 classes
Dense Built-Up (1)	Dense Built-Up (1)
Sparse Built-Up (2)	Sparse Built-Up (2)
Specialized Built-Up Areas (3)	Specialized Built-Up Areas (3)
Specialized Vegetative Areas (4)	Specialized Vegetative Areas (4)
Large Scale Networks (5)	Large Scale Networks (5)
Arable Lands (6)	Arable Lands (6)
Vineyards (7)	Vineyards and Orchards (7)
Orchards (8)	
Grasslands (9)	Grasslands (8)
Groces, Hedges (10)	Forests and semi-natural areas (9)
Forests (11)	
Open Spaces, Mineral (12)	
Wetlands (13)	Water Surfaces (10)
Water Surfaces (14)	

6.2.2 Multitemporal and multimodal network

In this study, a multitemporal and multimodal network pre-trained on MultiSenGE was used (Fig. 6.1). It was initially pre-trained on 4 Sentinel-2 dates spaced at least 17 days apart and the first available Sentinel-1 date per patch [Wenger et al., 2023b].
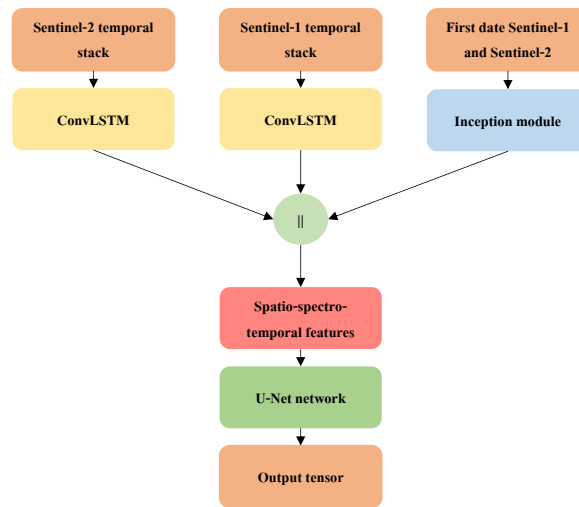
Figure 6.1 – *Pretrained multitemporal and multimodal network (ConvLSTM+Inception-S1S2 presented in Chapter 5)*

This method consists of a multitemporal and multimodal stack of the selected Sentinel-1 and Sentinel-2 patches as well as a monodate but multimodal stack of the first Sentinel-1 and Sentinel-2 dates of each patch. Then, a spatio-temporal feature extractor (ConvLSTM) is applied to both the Sentinel-1 multitemporal series and the Sentinel-2 multitemporal series. On the other hand, an Inception module, allowing the extraction of spatio-spectral features is applied this time on the monodata stack. The computed features are stacked and passed in a U-Net to compute a Land Use Land Cover map. The network thus pretrained on the MultiSenGE dataset can be applied and tested on other tiles and other cities in France.

6.2.3 Classification process

We chose to classify the set of each Sentinel-2 tile that we have previously segmented into patches of 256×256 pixels, the size of the input patches of the network. The areas of the cities studied are then cut out for each classification. Sentinel-2 images were downloaded according to several criteria, both for spatial and temporal inference:

- Less than 10% cloud cover and if no images are available on the desired dates, the lowest cloud cover available

- A complete image not containing no-data values (from the 290km swath of Sentinel-2)

- One image per month by maximizing the initial selection criteria used to train net-

work (17 days difference between two dates for the months of July, August, September and November)

Concerning Sentinel-1, we selected the first complete tile (still without no-data) preprocessed and sliced by the *s1tiling* processing chain, which does not change from the initial training method.

This classification process is applied for both temporal and spatial inference.

6.2.4 Temporal inference

To process temporal inference over Grand-Est region (Figure 6.2), we downloaded temporal series for 14 tiles over the study area. For 2018, the same dates as the initial pretrained model were chosen : July, August, September and November. For some months, it was not possible to match every image selection criteria (Section 6.2.3). If this case occured, we decided to extend the image search period by an additional month. Every date selected can be seen in the appendices (Appendices A.9 and A.10).
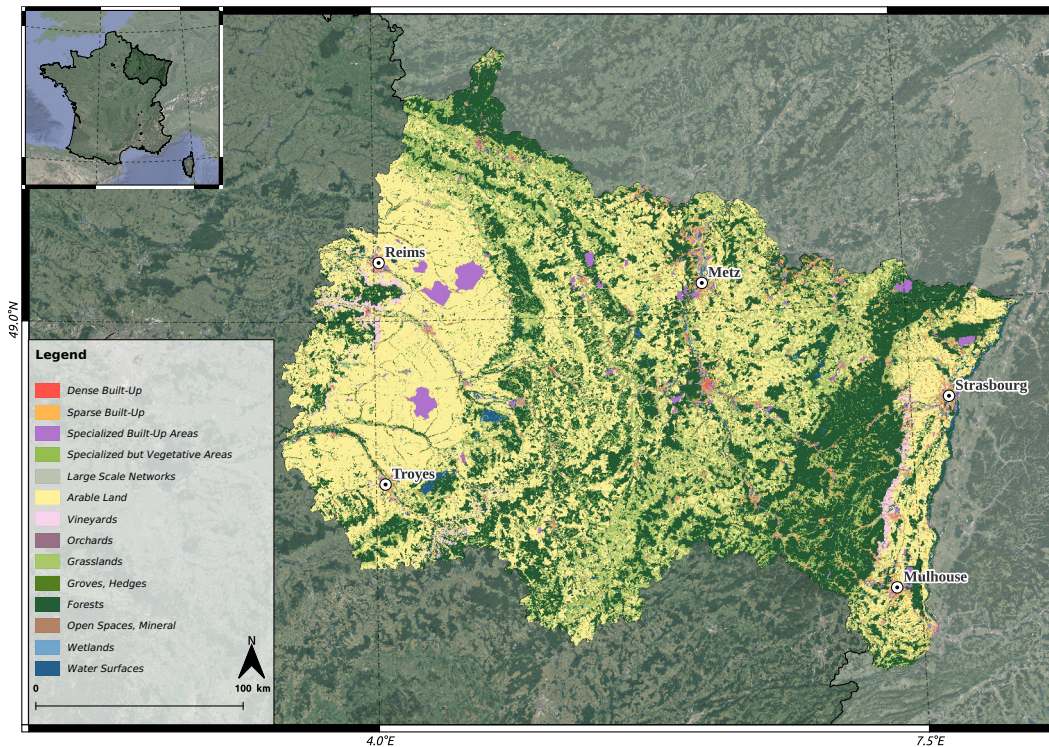

Figure 6.2 – *Reference data over Grand-Est for 2020.*

6.2.5 Spatial inference

Five cities morphologically similar to the cities of the Grand-East region were chosen to evaluate the generalization of a model trained on a region located several hundred of kilometers away : Toulouse, Dijon, Orleans, Lille and Rennes (Table 6.2 and Figure 6.3).

Table 6.2 – *French cities selected for classification*

City	Tile	Surface	Inhabitants
Toulouse	T31TCJ	118 km^2	400,000
Dijon	T31TFN	40 km^2	160,000
Orléans	T31UDP	27 km^2	116,000
Lille	T31UES	34 km^2	235,000
Rennes	T30UWU	50 km^2	220,000

Figure 6.3 – *Cities selected for spatial inference over France.*

In order to evaluate the generalizability of the model, three other cities were chosen, two in Western Europe (Bremen in Germany and Seville in Spain) and one in North Africa (Oran in Algeria), located in a much more desertic area with a urban morphology very different from Western cities (Figure 6.4).

Figure 6.4 – *Cities selected for spatial inference over Europe and North Africa.*

6.2.6 Classification evaluation

Temporal inference

As temporal inference is used to assess urban fabric changes between 2018 and 2022, we chose to perform visual inspection through both classifications over six major cities in the Grand-Est region : Châlons-en-Champagne, Metz, Mulhouse, Nancy, Reims and Strasbourg. In order to quantify and simplify changes interpretation, Sankey plots and maps has been made according to Appendices A.12 and A.13. Only changes from natural areas to UF has been ploted as classification errors has been discovered for 2022 for these surfaces probably due to the date selection criteria (explained in section 6.2.4). Spectral response of natural areas are not the same between April and August than between June and November, epecially with a year 2022 particularly dry (Appendices A.20, A.21, A.22, A.23, A.24, A.25). As we wanted to intentify abrupt changes in urban fabric, missclafications for natural areas were not a limiting factor for the objectives of this study.

Spatial inference

The evaluation of the spatial inference classifications was primarily based on a qualitative assessment, as similar baseline data was not available at the selected study sites. Nevertheless, in order to provide a qualitative assessment essential to any classification work, a digitalization of the five selected urban classes was performed for each of the five cities (Appendices A.14, A.15, A.16, A.17, A.18, A.19). The sixth class is represented as the concatenation of the five others natural classes. We have done this digitalization using a Sentinel-2 image as a background. Also, Urban Atlas[4] was used for each city as a decision support to discriminate very complex areas. For this process, we digitized five areas of 4 km^2 each for each city starting from the urban centre towards the suburbs.

To perform qualitative assesment for spatial inference, we chose to use Recall, Precision and $F1_{Score}$ (also know as Dice) metrics for each city and for each class :

$$Recall = \frac{TP}{TP + FN} \tag{6.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{6.2}$$

$$
\begin{aligned}
F1_{Score} &= \frac{2 \times Precision \times Recall}{Precision + Recall} \\
&= \frac{2 \times TP}{2 \times TP + FP + FN}
\end{aligned}
\tag{6.3}
$$

where TP represents the True Positive, FP the False Positive and FN the False Negatives. Weighted $F1_{Score}$ is calculated with the mean of each class according to each class support.

6.3 Results

6.3.1 Spatial inference

Quantitative assessments

Weighted $F1_{Score}$ classification metric was computed for each city over the five subset areas (Table 6.3). We also performed $F1_{Score}$ for each class in Table 6.4. As a reminder, class (6) is the concatenation of all natural classes (class (6) to (10) in Table 6.1). As shown in Table 6.3, Dijon is the city where the weighted $F1_{Score}$ is the higher with a score of 0.8087 following by Orléans with 0.7742. Lille has the least score with 0.6866.

---

[4] https://land.copernicus.eu/local/urban-atlas/urban-atlas-2018

Table 6.3 – *Weighted $F1_{Score}$ for each city.*

Cities	Score
Toulouse	0.7029
Dijon	**0.8087**
Orléans	0.7742
Lille	0.6866
Rennes	0.7187

Concerning each class, we clearly see that Dijon has the best $F1_{Score}$ per class for 4 out of 6 classes. Also, it outperforms the other cities for class (1), (3) and (5) as seen in Table 6.3. As presented in Chapter 5, class (4) remains complex to classify, even in dense urban areas. The same observation can be applied for class (5). Large Scale Networks are complex to detect in cities as they are often less then one pixel wide.

Table 6.4 – *$F1_{Score}$ per class for each city.*

Cities selected	Classes					
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Toulouse	0.7137	0.7255	0.7021	**0.6072**	0.3855	0.7819
Dijon	**0.8005**	0.8140	**0.7774**	0.5363	**0.6154**	**0.8996**
Orléans	0.6408	**0.8462**	0.6989	0.4215	0.4885	0.8355
Lille	0.6315	0.7864	0.6360	0.5835	0.4930	0.8082
Rennes	0.6446	0.6892	0.7012	0.5209	0.5666	0.8589

Qualitative assessments

Land cover maps presented in Figure 6.5 confirmed encouraging classification results. Salt and pepper classification noise, which was often detected for pixel approaches (e.g. Random Forest) and especially for urban fabric mapping, is almost none for these classifications. For each city, urban city centre is well identified and correspond to Dense Built-Up for western cities.
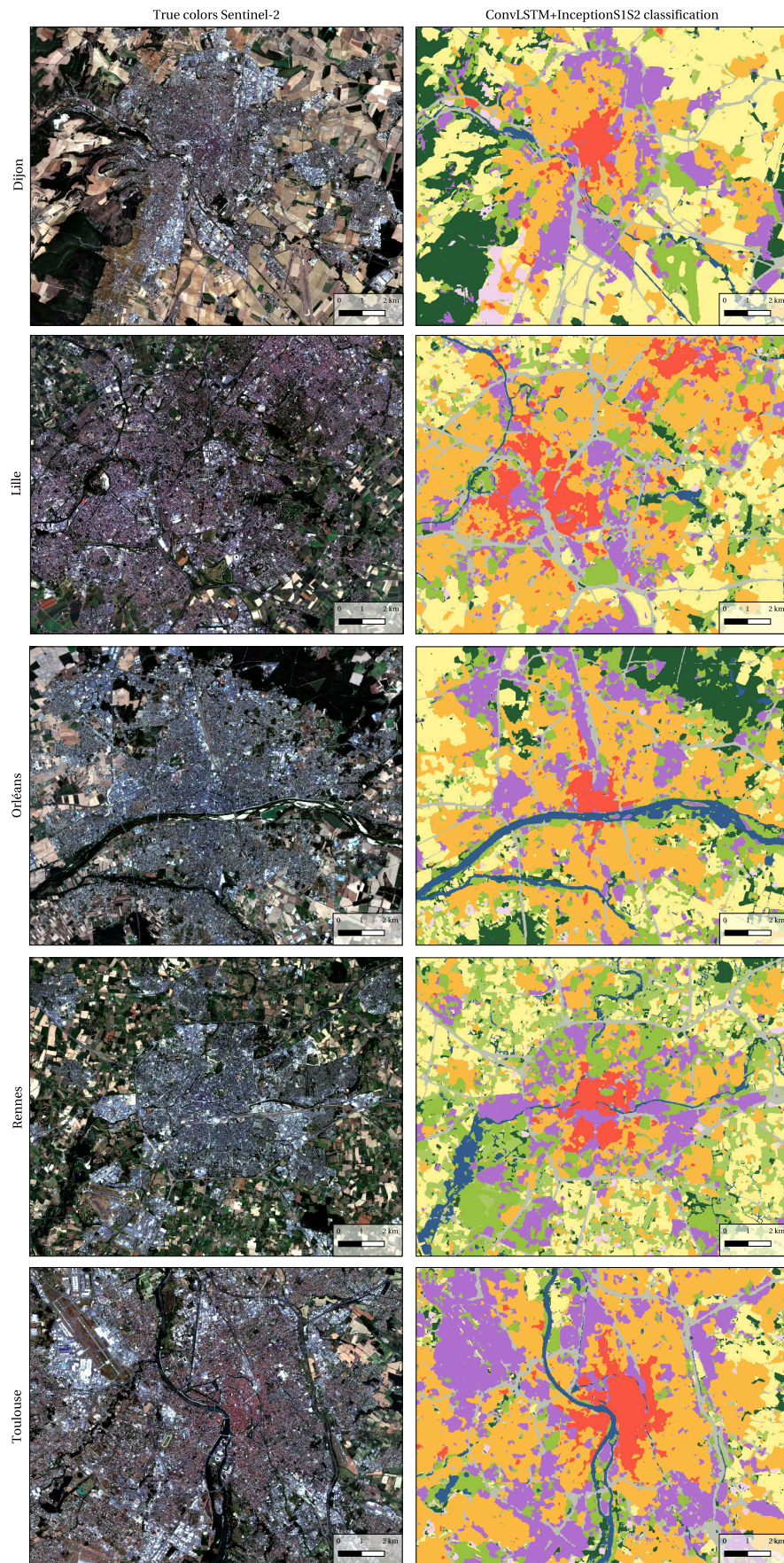
True colors Sentinel-2   ConvLSTM+InceptionS1S2 classification

Dijon

Lille

Orléans

Rennes

Toulouse

Figure 6.5 – *Classification results for the five cities studied.*

As seen in Figure 6.5, urban boundaries are well classified on almost each city. The most important confusion is with class (4), which represents Specialized but Vegetative areas (Table 6.1). These areas are correctly classified when they are inside dense urban areas but, at the boundaries of the city, we can relate some confusion with natural areas, especially for Orléans and Rennes. More results are presented in Appendices A.14, A.15, A.16, A.17, A.18, A.19.
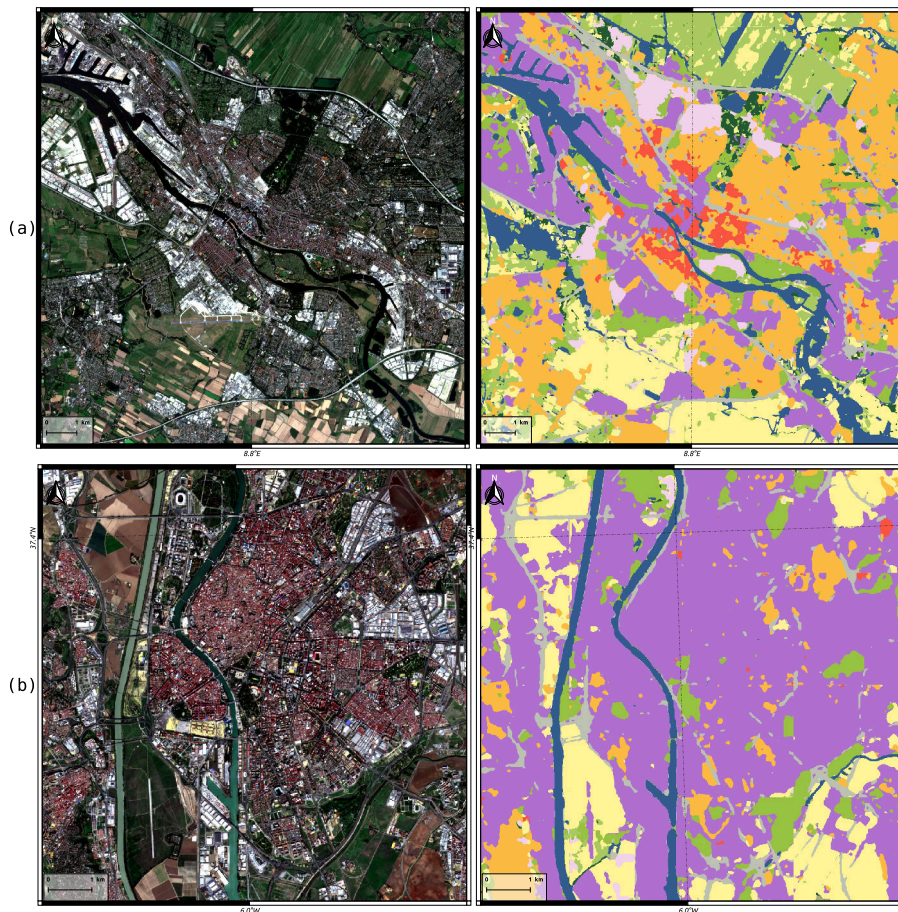

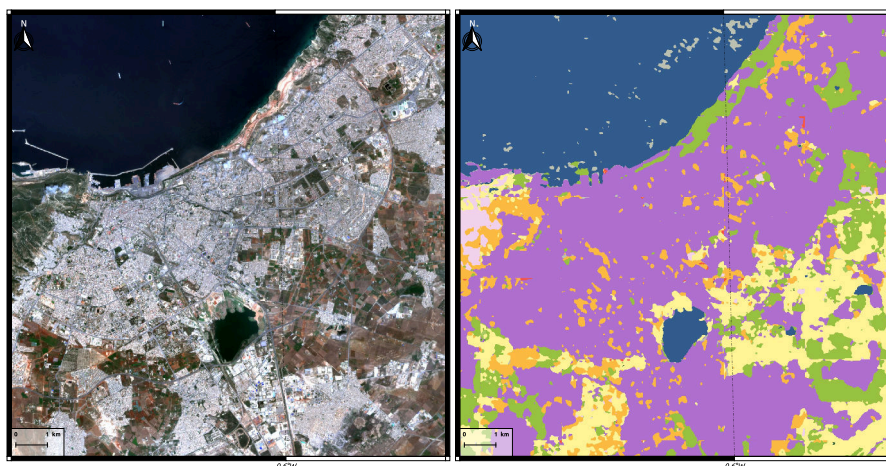Figure 6.6 – *Classification results for (a) Bremen (Germany) and (b) Seville (Spain).*

Figure 6.7 – *Classification results for Oran (Algeria).*

In opposition to the French cities, results of the European and African cities are contrasted. We notice visually consistent results for the city of Bremen in Germany, while for Seville (Spain) and Oran (Algeria), a strong confusion is found between the classes, both for the urban fabrics and the natural classes (Figure 6.6 and 6.7). In fact, for urban fabric, *Specialized Built-Up Areas (3)* is the majority and classifies the entirety of the city. On the other hand, *Specialized but Vegetative Areas (4)* are quite well classified with few false positives except for Oran where the cimate is very arid. The dataset is part of a semi-continental climate with a building morphology typical of northern European cities. The Mediterranean climate being characterized by higher temperatures and a drier landscape than northern France and Europe. One of the hypotheses of the misclassification for Seville and Oran may be due to a drier climate and a lower presence of chlorophilic vegetation, both within the urban fabric and in natural areas.

6.3.2 Temporal inference and change detection

Performing temporal inference at a large scale (Grand-Est region) for two different dates (2018 and 2022) enabled us to detect, after maps comparison, brutal changes in the urban areas (i.e. roads, residential buildings, residential areas ...). As seen in Figures 6.8, 6.9, 6.10, 6.11, 6.12 and 6.13, we can extract urban changes at the cost of natural areas. It is also possible to have the information about the changing class, such as in Figure 6.10 where natural areas are turning *Specialized Built-Up Areas (4)* and *Sparse Built-Up (2)* areas between 2018 and 2022. This method allows even the smallest changes in the territory to be detected, such as small housing blocks (Figure 6.12). Construction sites or recent housing developments where the reflectance is quite high (Figure 6.9), are nevertheless detected as *Specialized Built-Up Areas (3)*.

However, we can still detect false positives on every subset especially for *Specialized But Vegetative Areas (4)*. This is not an incovenient as these change detection maps should be analyzed by expert. They will detect changes and can keep real LULC changes.
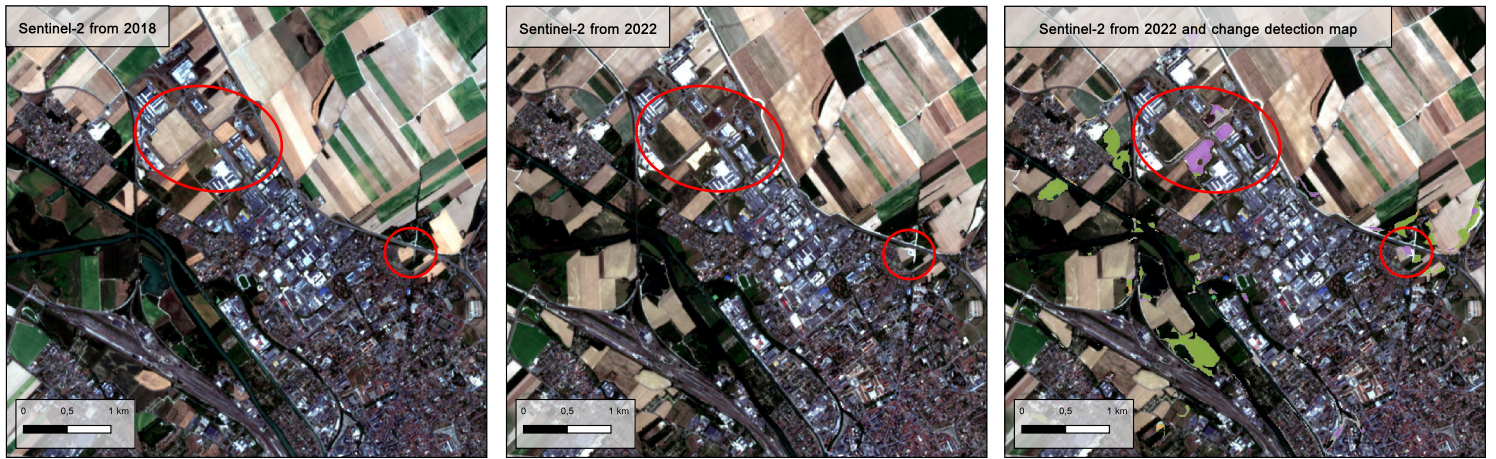
Figure 6.8 – *Change detection near Châlons-en-Champagne (subset 3 maps the class to which the surfaces have changed).*
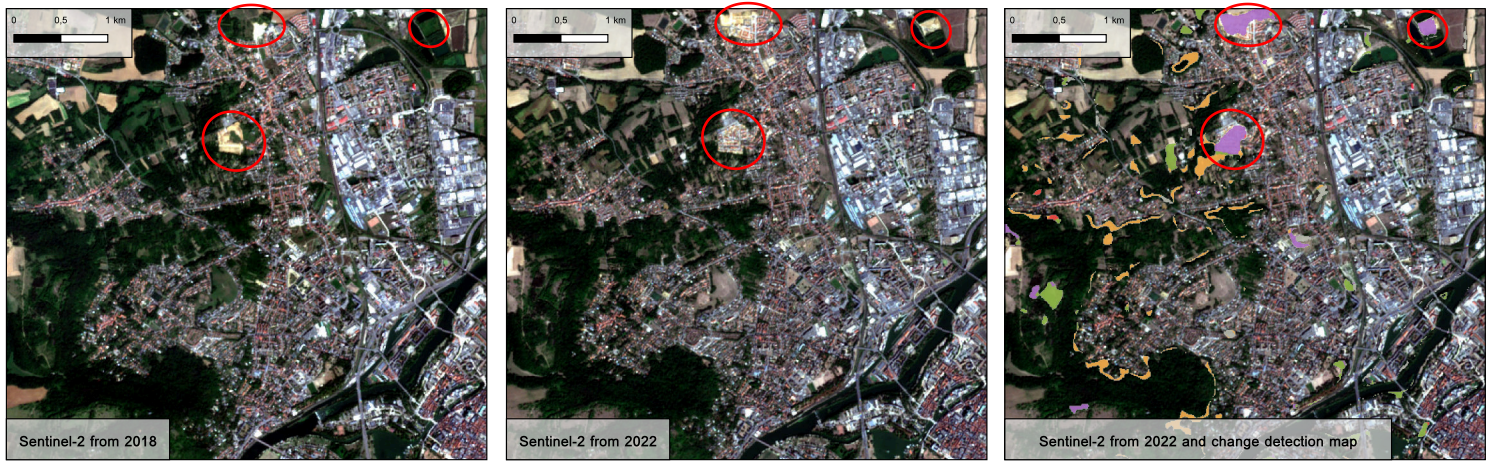

Figure 6.9 – *Change detection near Metz (subset 3 maps the class to which the surfaces have changed).*
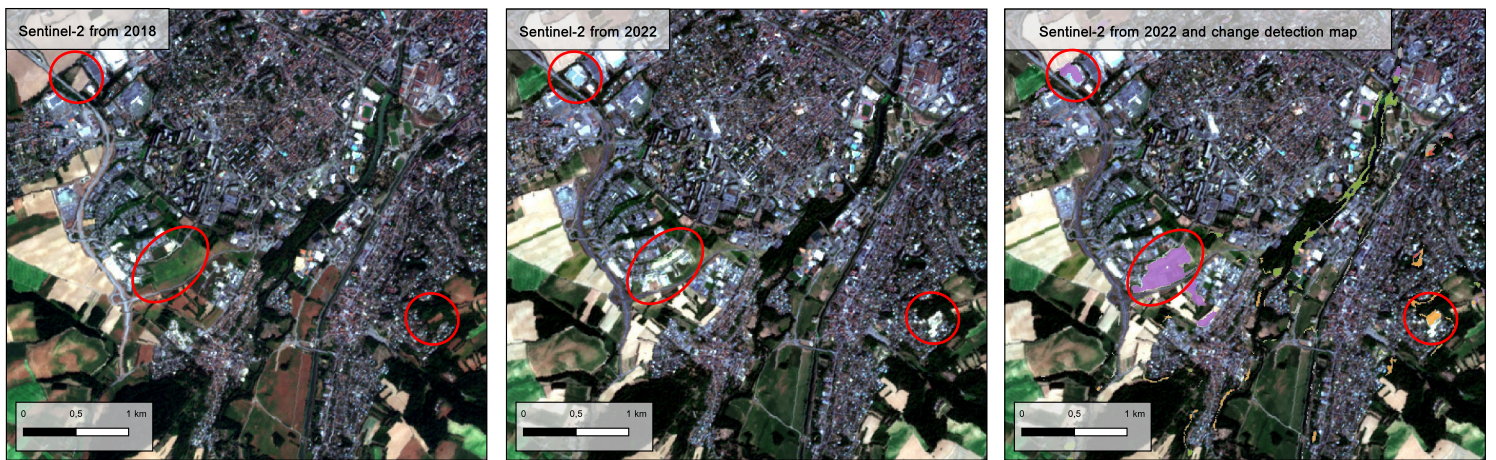

Figure 6.10 – *Change detection near Mulhouse (subset 3 maps the class to which the surfaces have changed).*

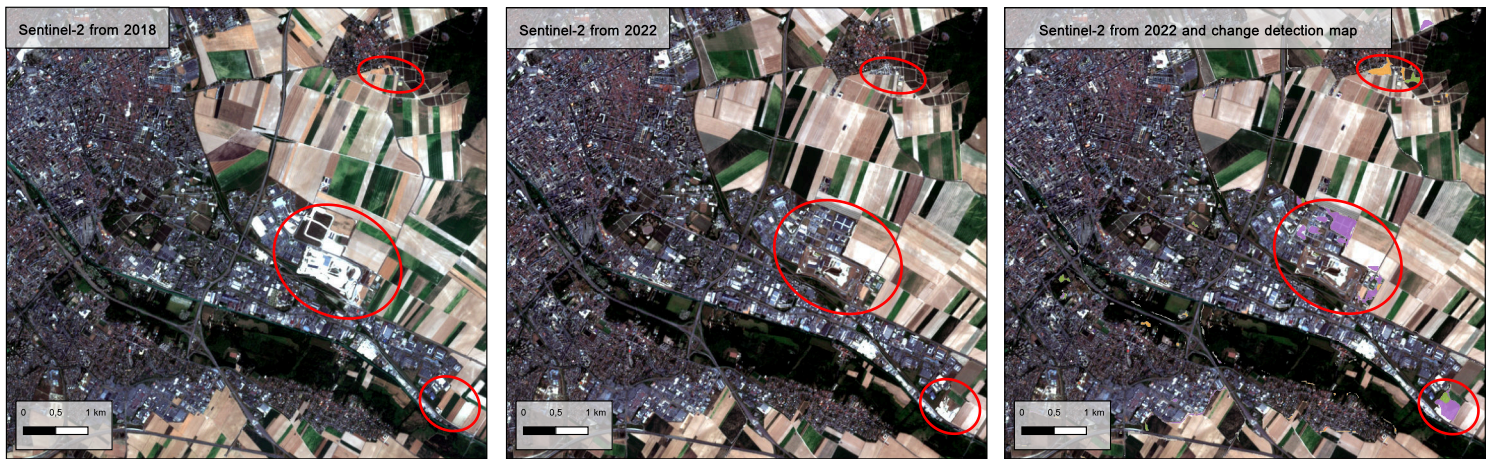Figure 6.11 – *Change detection near Nancy (subset 3 maps the class to which the surfaces have changed).*

Figure 6.12 – *Change detection near Reims (subset 3 maps the class to which the surfaces have changed).*
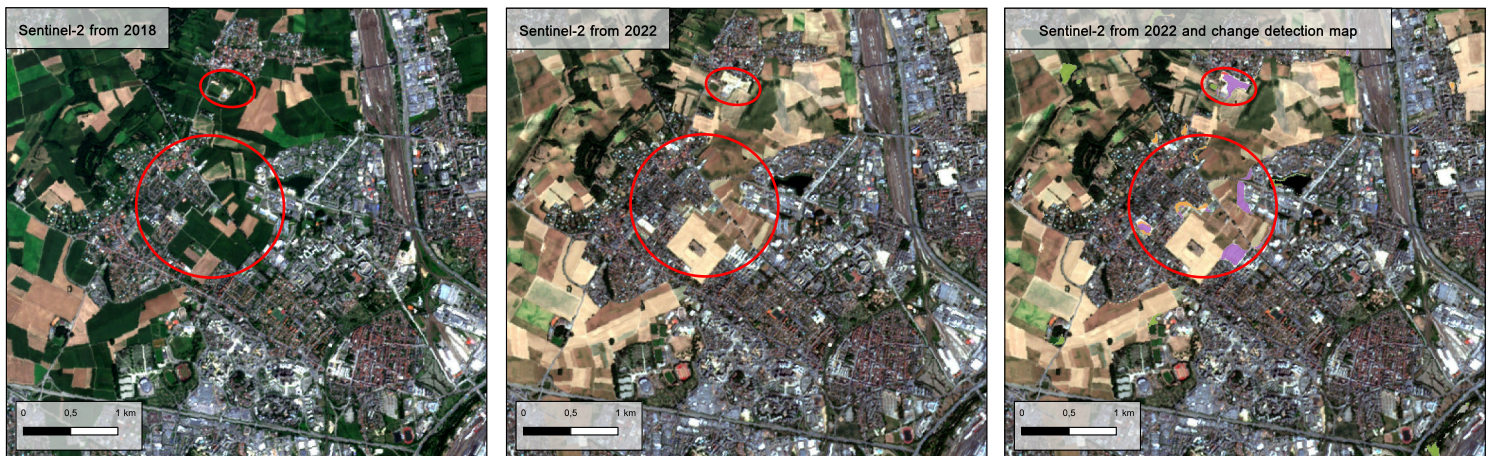

Figure 6.13 – *Change detection near Strasbourg (subset 3 maps the class to which the surfaces have changed).*

Through Figures 6.14 and 6.15, we can see the LULC change for each subset for UF. Unfortunately, due to the false positive, it is hard to quantify changes before expert interpretation. Also, we can identify than the major changes are between *Arable Lands (6)*

and *Specilized Built-Up Areas (3)* for most of the subsets. Both Grand-Est classifications (2018 and 2022) can be seen through a webmap[5]. Due to the numerous confusions between classes, especially at the edge of urban areas, the Sankey diagram contains several errors, especially for the evolution of Water Surfaces towards UF which is not a class sensitive to artificialization. These Sankey diagrams allow, even if many false positives are also present, to obtain a global estimate of the change surfaces. This is the case for Strasbourg (Figure 6.15 (c)), where two large changes are known and identified: the construction of a shopping centre (*Shopping Promenade*) and the construction of a highway network (*Grand Contournement Ouest - GCO*). There are indeed significant changes in the areas for the *Large Scale networks (5)* and *Specialized Built-Up Areas (3)* classes. It is mainly the *Arable Land (6)* that has changed to these two UF classes.

---

[5]https://romainwenger.fr/multisenge/Grand_Est.html
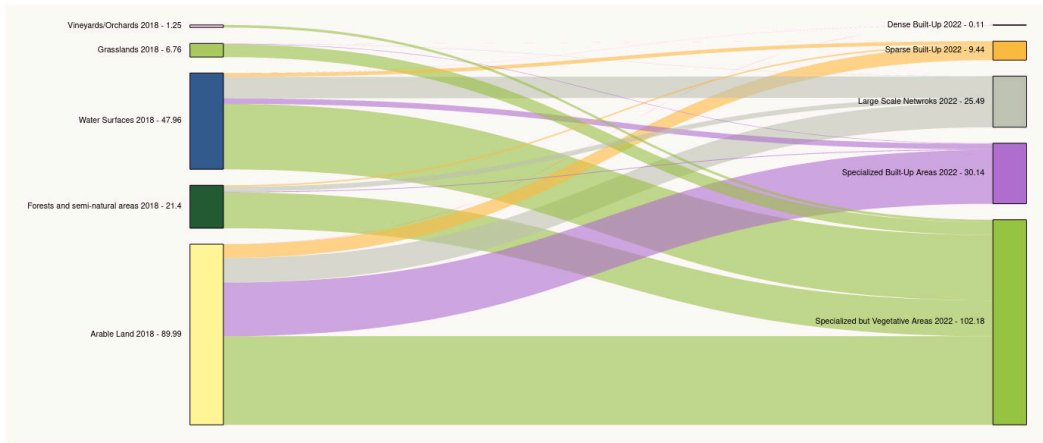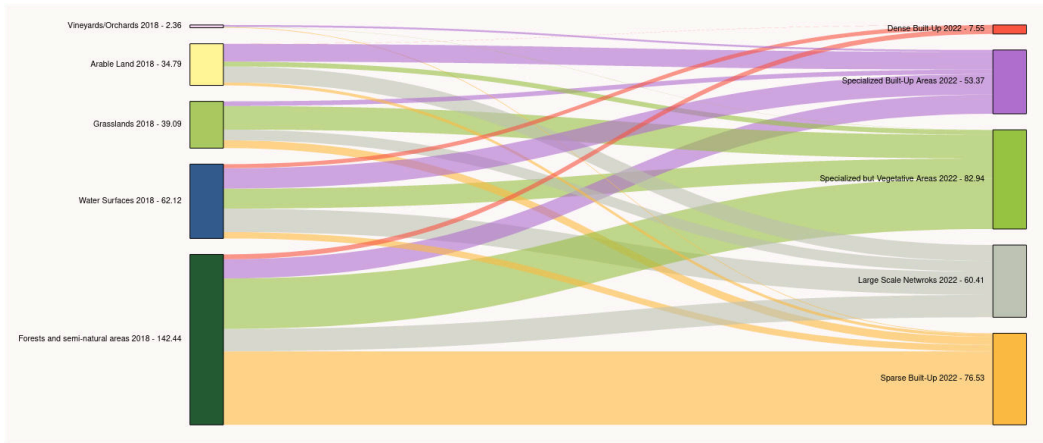
(a) Sankey plot for Châlons-en-Champagne
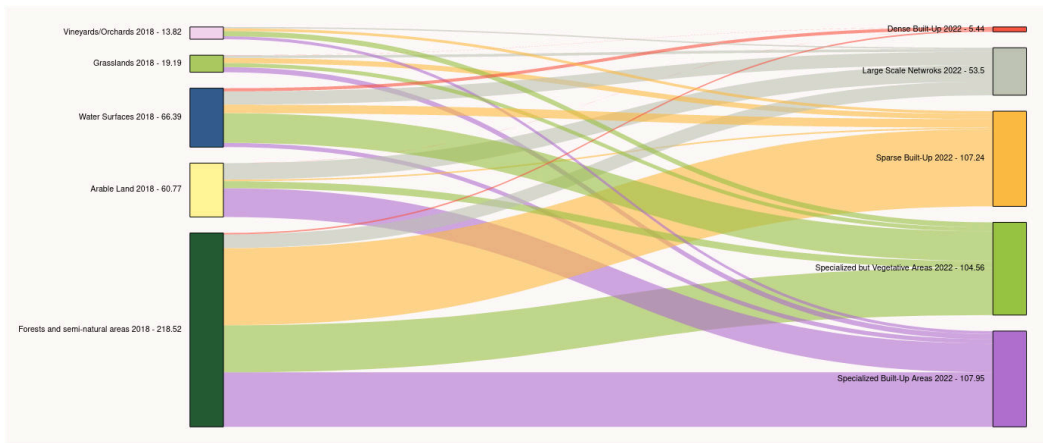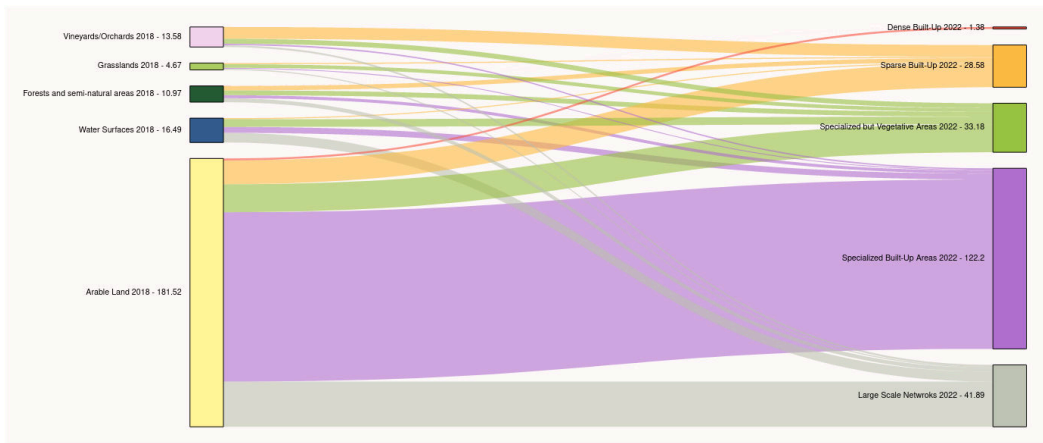

(b) Sankey plot for Metz
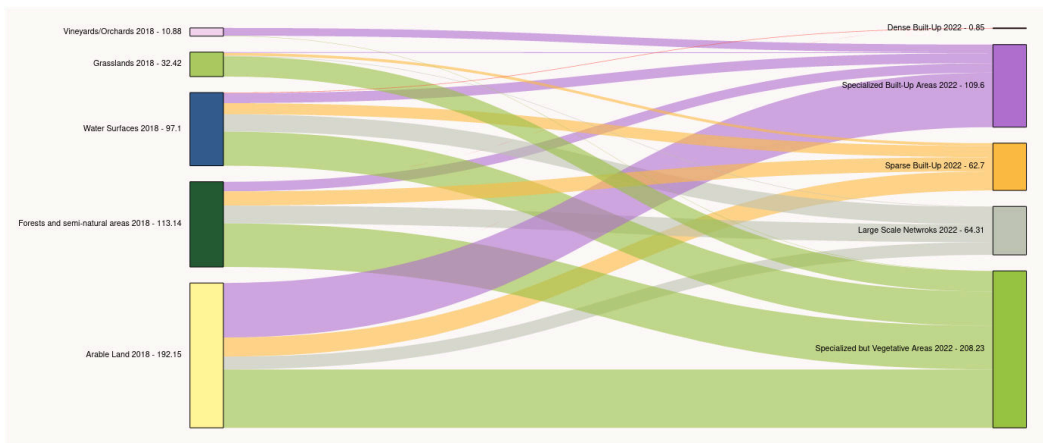
(c) Sankey plot for Mulhouse

Figure 6.14 – *Changes from natural areas to urban fabric between 2018 and 2022 for six cities over Grand-Est region.*
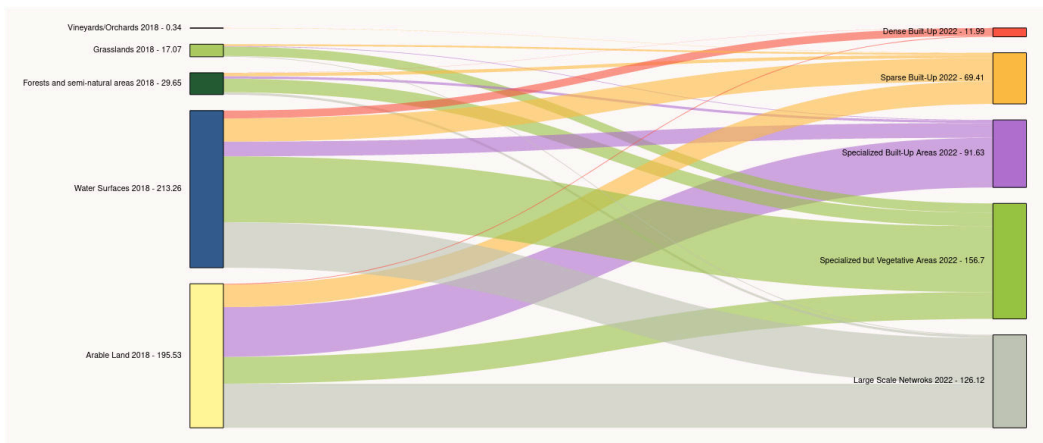
(a) Sankey plot for Nancy

(b) Sankey plot for Reims

(c) Sankey plot for Strasbourg

Figure 6.15 – *Cont.*

6.4 Conclusion

This chapter has shown the capacity of a deep learning model, trained over a specific french region, to generalize over others cities in France. The model trained over Grand-Est region in France (*MultiSenGE* dataset) was applied for five cities in France, two in Europe and one in North Africa to evaluate the genericity of a LULC multitemporal and multimodal semantic segmentation network. We also applied temporal inference between two dates (2018 and 2022) for the same region as the training dataset to monitor LULC dynamics. With this study, we answer to the third objective which was to assess spatial and temporal inference capacity of the proposed classification models to generalize the proposed approach for mapping large-scale territories and analyzing their dynamics.

Results are encouraging and illustrate that models trained on this dataset can achieve great performance over other western geographical areas for UF mapping. Thus, it would not be mandatory to perform transfer learning or domain adaptation for UF mapping over France. However, a small convergence of the weights, as a results of a new training phase over local ground reference data, could improve current results. Temporal inference has shown the capacity of a deep learning model to perform change detection between two classifications at two different dates. This open several perspectives, especially for urban planners as they will be able to assess LULC changes through years easilly. These products can be used by urban planners to improve existing manual classifications at low cost (financial and time). At this resolution, these results can be used by end-users as a warning system for change detection.

GENERAL CONCLUSION AND PERSPECTIVES

<div style="text-align: right; font-size: 3em;">7</div>

CONTENTS

7.1 Summary and conclusion of the work

Through this thesis, the objective was to assess the potential of optical and SAR high spatial resolution time series for mapping Land Use Land Cover (especially for urban fabrics) and for analyzing its dynamics at large scale to help end-users to obtain up-to-date information for their urban applications.

Methodological issues have been focused on the development and evaluation of deep learning methods for (1) the automatic mapping of LULC and urban fabrics by using multitemporal and multimodal Sentinel-1&2 imagery and (2) the generalisation of our approach (spatial inference) and the monitoring of LULC dynamics (temporal inference). Thematic issues of this thesis were mainly focused on urban fabric where few automatic classifications focus on discriminating these areas into more than four classes from Sentinel images.

The key findings and the main contributions of this PhD research work can then summarized into four points:

1. The **construction of an innovative dataset** in order to propose to the research community an original dataset with multimodal, multitemporal imagery and reference data adapted to 10m based on Sentinel-1&2 covering a large area (the Grand est region) representing 1/7 of the French territory. This dataset allowed us to perform all the experiments of this thesis.

2. We contributed to the formalization of the use of **deep learning methods for UF mapping**. Results from the processing of mono-temporal optical imagery have shown an interest in using these semantic segmentation approaches to provide up-to-date mapping of urban fabric, which is often not available with HSR (High Spatial Resolution) optical imagery. The feature fusion between hexogenes and spectral band indices showed a gain and a better performance for the classification of these surfaces that exceeded the current results and allowed the addition of one more semantic class, essential for the classification of urban areas. These approaches have therefore demonstrated the value of deep learning methods for classifying urban fabric from HSR imagery (sub-objective 1).

3. The **integration of multitemporal optical and SAR imagery** has improved the results from single-time imagery. The results were convincing and showed the influence of multimodal and multitemporal data and a semantic consistency for urban and rural areas with a strong reduction of confusion compared to optical mono-

temporal approaches. Furthermore, even at 10m spatial resolution, intra-annual vegetation dynamics have an influence in UF classification (sub-objective 2).

4. The generalization of **approaches with spatial and temporal inference** has shown through the spatial inference approach that results trained on a region with a semantic segmentation network is able to generalize to other regions which must still be quite similar semantically to the original region (e.g. this method seems not suitable for north african cities). This makes it possible to propose a fast and coherent map without having to re-train a deep learning network over a specific region. Temporal inference offers end-users rapid change detection from a pre-trained network. It allows to detect changes between two different dates from the training date of the model (sub-objective 3).

7.2 RESEARCH PERSPECTIVES

Improving the MultiSenGE reference data. The current reference data contains only nine natural classes. Even if it is semantically adapted to the 10m spatial resolution imagery, it remains rather restrictive, since it is possible to distinguish natural areas, and in particular arable land, in many more classes [Inglada et al., 2017, Rußwurm et al., 2019]. So, a first perspective would be to increase the number of semantic classes focused on natural areas. A second perspective would be to improve the dataset by adding all the optical satellite images without restriction on the cloud cover. In the short term, we plan to add the whole optical time series with a cloud cover mask per patch. It could be also possible to enrich the dataset with multi-source data with fine spatial resolution (e.g. with Pleiade and/or SPOT 6/7 VHSR image by patch).

Combination of HSR and VHSR. The combination of HSR and VHSR images is one of the key issues in the application of semantic segmentation using deep learning methods. Indeed, the geometry is different due to the spatial resolution and the acquisition mode, which makes the feature fusion in a network more complex than between geometrically similar images. The study of this combination is one of the current challenges, because VHSR images, although not available as time series, bring a refinement to the classification by improving the detection and the discrimination between semantic classes. One of the existing approach, $MMCNN_{SD}$ [Gbodjo et Ienco, 2021], is to train several classifiers in parallel and face them with a combinatorial loss to perform an auto-distillation, which consists in confronting the different classifiers to find the most adapted weights in the same network. By analogy, this is very similar to knowledge distillation (transfer

of knowledge from a teacher network to a student network). To group all the features, the method presented in Figure 7.1 sums the features of the three inputs (Sentinel-1 SITS, Sentinel-2 SITS and SPOT panchromatic and multispectral) and then compares this "main" loss with the auxiliary classifiers. This is an interesting strategy and we would also try to add multi-source data in our model for UF mapping. This perspective is part of the continuity of the thesis work, which will be performed as a post-doctorate research at the Image Ville Environnement (CNRS-UMR 7362) laboratory.
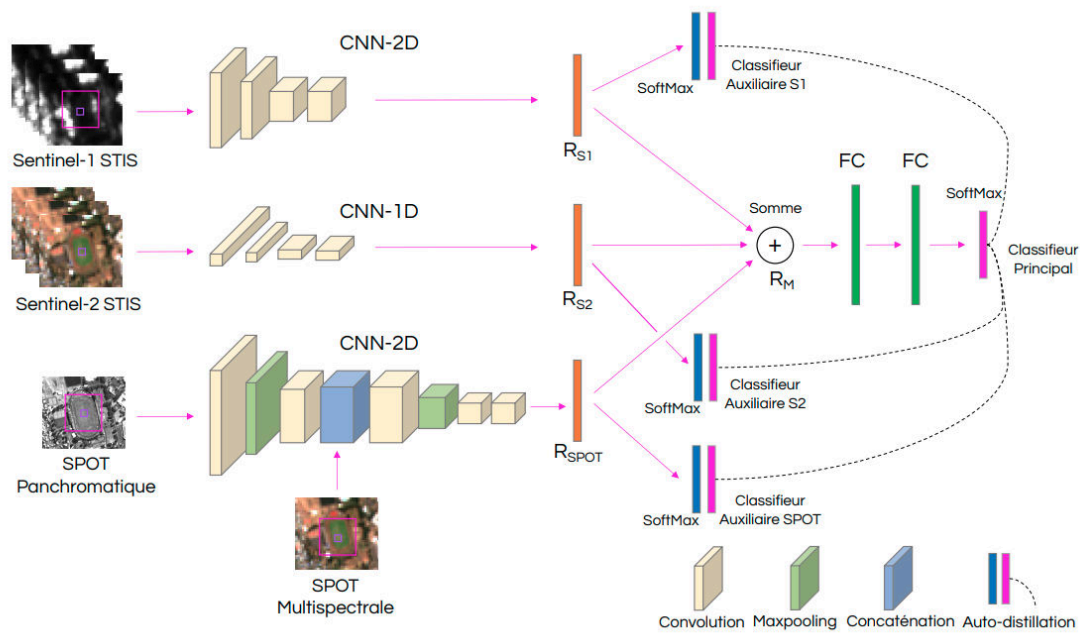

Figure 7.1 – *MMCNN$_{SD}$ framework using multimodal and multitemporal satellite imagery for Land Use Land Cover applications (methodology developped by Gbodjo et Ienco [2021]).*

Towards a large scale LULC and UF product for end-users. Another perspective would be to propose an improvement of the OSO product by integrating our classification results focused on UF in their existing LC product. Indeed, the *ConvLSTM+Inception-S1S2* model developed in this thesis is suitable for multi-temporal and multimodal data fusion and urban fabric mapping. The generalization of this method with a small amount of training data opens new perspectives for large-scale mapping. By applying transfer learning methods and using the weights calibrated for the Grand Est region, the model can be easily trained for typologically different regions. Some research would be necessary to propose an optimal integration of two products by using fusion methods for instance. However, this type of processing is challenging in the remote sensing field due to the multiplication of existing LULC productions. A last perspective would be to adapt our network to South cities in order to discriminate UF by focusing on the distinction between formal and informal settlement.

RÉSUMÉ ÉTENDU EN FRANÇAIS

8

CONTENTS

8.1 Introduction

La couverture biophysique de la surface terrestre, autrement dit l'occupation des sols, fait partie intégrante depuis plusieurs années des Variables Climatiques Essentielles (VCE). Dans un contexte de changement global, la connaissance précise de l'occupation des sols est une composante importante de nombreuses recherches en environnement. En effet, celle-ci est une donnée essentielle en entrée de plusieurs types de modèles (par exemple climatique, hydrologique) et elle peut aider les acteurs locaux dans les prises de décision. Les cartes d'occupation des sols, principalement produites par interprétation visuelle et digitalisation de photographies aériennes ou d'images satellites, nécessitent souvent plusieurs mois, voire années de production sur de grands territoires et avant diffusion au public. Depuis plusieurs années, les constellations de satellites d'observation de la Terre sont capables de transmettre des images à haute résolution spatiale et temporelle suivant plusieurs modes d'acquisition (radar, optique …). Certaines missions satellitaires, telles que les missions Sentinel à travers le programme Copernicus, fonctionnent en constellation permettant d'obtenir plus d'une image par semaine du même endroit de la Terre. Cette haute fréquence temporelle permet de quantifier des dynamique anthropiques et naturelles pour suivre avec précision l'évolution des territoires. Avec leur haute résolution spatiale, les images Sentinel offrent également une résolution adaptée pour la cartographie de l'occupation des sols sur l'ensemble du territoire. Par exemple, l'un des produits d'occupation du sol à l'échelle national (OSO du Pôle Theia), est produit automatiquement chaque année pour la France entière et quelques pays frontaliers à partir des images Sentinel-2. L'avènement des méthodes d'intelligence artificielle a permis à la communauté scientifique en télédétection l'automatisation de la cartographie de l'occupation des sols avec des résultats très prometteurs. Les premiers travaux dans ce domaine utilisaient des méthodes supervisées dites « classiques », telles que les arbres de décision (Random Forest), les machines à vecteurs de support (Support Vector Machine – SVM) ou encore les K-moyennes (K-Means – pour les approches non-supervisées). Ces méthodes ont récemment été surpassées par les approches d'apprentissage profonds et plus particulièrement de segmentation sémantique, notamment depuis l'évolution des méthodes de calcul et le calcul sur carte graphique. Celles-ci nécessitent des données massives pour produire des résultats exploitables et ont pour but d'être généralisées à grande échelle.

Ainsi, il est nécessaire d'avoir des carte d'occupation des sols à jour et précise pour une variété d'application telles que le suivi de l'environnement, l'analyse des risques ou encore l'aménagement des villes. Le problème général de cette thèse concerne la cartographie de l'occupation des sols mais plus particulièrement les tissus urbains dans un contexte de *big*

data où les données collectées quotidiennement dépassent le petaoctet, comme par exemple celles issues des capteurs Sentinel-1&2.

Dans ce contexte, l'objectif principal de cette thèse est d'évaluer la contribution de l'imagerie spatiale mulitemporelle et multimodale (Optique et Radar) pour la classification des modes d'occupation des sols et en particulier les morpho-types urbains appelés aussi tissus urbains. Il s'agit plus précisément : (1) d'évaluer la performance des méthodes de classification fondée sur l'apprentissage profond, (2) d'évaluer l'apport de l'imagerie spatiale multitemporelle et multimodale fournies par la constellation de satellite Sentinel-1&2 et (3) d'évaluer la capacité d'inférence spatiale et temporelle des modèles de classification proposés. Ces travaux se placent dans un contexte de traitement massif de données, à large échelle (sur de grands territoires géographiques).

Ce résumé long de thèse est organisé en six sections en incluant l'introduction où pour les trois premières correspondent à trois articles scientifiques publiés dans le cadre du doctorat.

- La seconde section se concentre sur la zone d'étude et la création du jeu de données MultiSenGE. L'objectif de ce chapitre est de préparer une donnée de référence sur l'ensemble de la zone d'étude [Wenger et al., 2022a], ce qui est un *challenge* important dans la communauté en télédétection.

- La troisième section cherche à évaluer le potentiel de l'imagerie mono-temporelle optique Sentinel-2 pour la classification des tissus urbains à partir d'approches de segmentation sémantique et de méthode de fusion de caractéristiques [Wenger et al., 2022c].

- La quatrième section porte sur la combinaison de données Sentinel-1&2 (multitemporelles et multimodales en utilisant le jeu de données MultiSenGE) pour la cartographie automatique des tissus urbains [Wenger et al., 2023b]. L'objectif est d'évaluer la contribution de chaque capteur et des méthodes de fusion de caractéristiques par rapport à l'imagerie mono-temporelle optique.

- La dernière section explore l'inférence spatiale et temporelle dans un contexte de reproductibilité des méthodes multitemporelles et multimodales développées dans la section précédente. L'objectif de cette section est double : aller vers une cartographie de l'occupation des sols à grande échelle à partir des méthodes d'inférence spatiale et détecter des changements brusques (par exemple l'étalement urbain) avec les méthodes d'inférence temporelle.

8.2 Zone d'étude et jeu de données MultiSenGE

Le développement constant des missions d'observation de la Terre a permis l'acquisition d'une grande quantité de données satellites qui ont considérablement changé la manière dont l'humanité gère son territoire. Les missions Sentinel-1 et Sentinel-2, du programme Copernicus développé par l'ESA, permettent d'obtenir des images radar et multispectrales avec une fréquence de revisite très courte. Ces images sont utiles pour étudier la dynamique des processus et des objets d'intérêt, en particulier pour la classification d'occupation du sol.

A notre connaissance, aucun jeu de données utilisant de l'imagerie Sentinel-1&2 et une description des tissus urbains en 4 à 5 places n'était disponible. Pour remédier à cela, nous avons créé le jeu de données MultiSenGE qui offre des triplets de données Sentinel-1, Sentinel-2 et d'occupation du sol pour la région Grand-Est en France, avec une précision géométrique élevée. Ce jeu de données est destiné à être utilisé pour la segmentation et la classification sémantique des surfaces urbaines en 5 classes d'occupation du sol. Ce jeu de données est comparé à d'autres jeux de données existants, tels que BigEarthNet [Sumbul et al., 2021] ou SEN12MS [Schmitt et al., 2019a], qui utilisent des données de référence CORINE et MODIS Land Cover, respectivement. MultiSenGE offre ainsi un nouveau jeu de données pour la recherche en télédétection.

MultiSenGE couvre l'ensemble des 14 tuiles Sentinel-2 de la région Grand-Est (Figure 8.1), l'une des plus grandes régions de France avec ses 57,433 km^2. Cette région a été choisi grâce à la mise à disposition d'une donnée d'occupation des sols de référence, la donnée OCSGE2-GEOGRANDEST qui consiste en une photo-interprétation de photogra-

phies aériennes de 2019/2020. L'avantage de cette donnée est sa grande précision avec des Unités Minimum de Collectes (UMC) de $50m^2$ pour les milieux urbains. Afin d'avoir une cohérence sémantique et spatiale avec la données satellitaire (10m de résolution spatiale), OCSGE2 a été rééchantillonnée pour obtenir une données de 14 classes d'occupation des sols. Une données annexe (BDTOPO-IGN) a été utilisé pour l'extraction des routes majoritaires (autoroutes, routes européennes et routes nationales) puis ajouter en fin de traitement à la données finale.
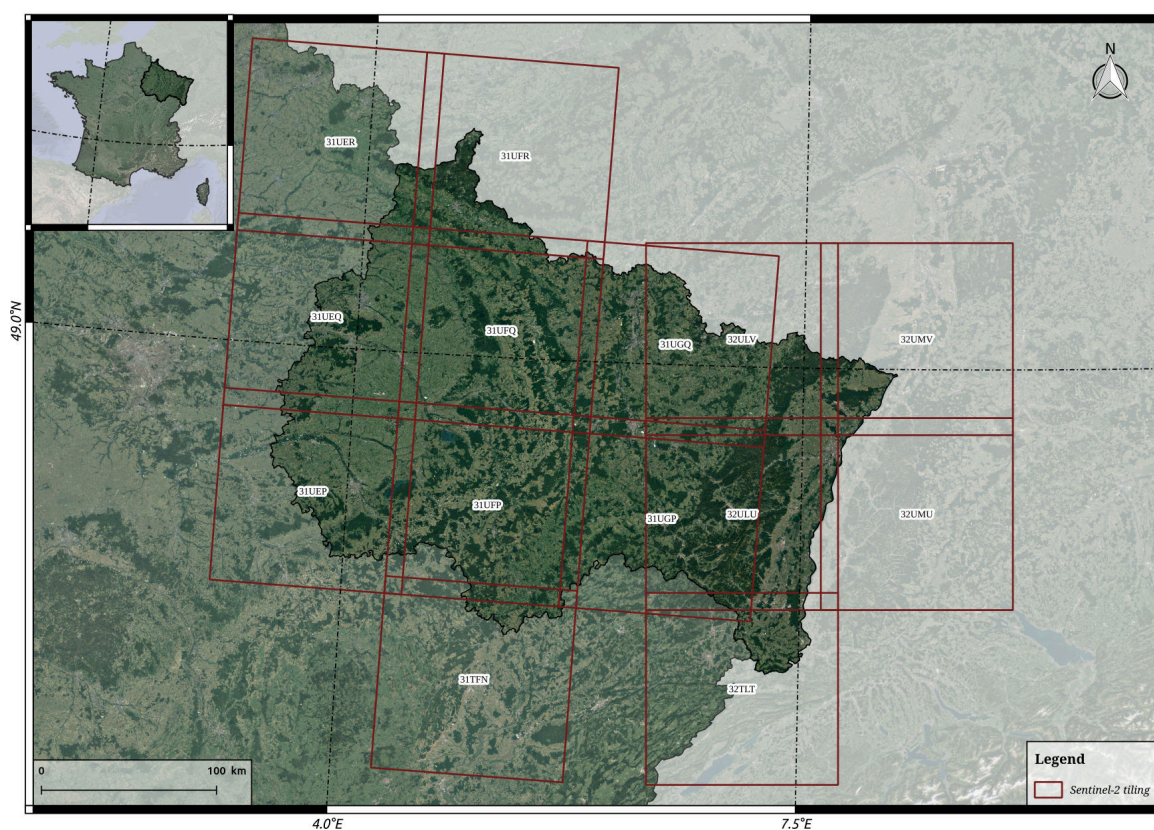

FIGURE 8.1 – *Zone d'étude avec le tuilage Sentinel-2.*

Les données satellitaires en entrée de MultiSenGE proviennent des capteurs multimodaux Sentinel-1&2 (radar et optique). La données Sentinel-1 a été téléchargée et prétraitée en utilisant la chaine de traitement *s1tiling* développée par le CNES (Centre Nationale des Etudes Spatiales) alors que la données Sentinel-2 a été téléchargée via le portail de téléchargement Théia en utilisant le script *theia_download* [Hagolle, 2021]. Les 10 bandes à haute résolution spatiale (10m et 20m) ont été conservé.

Pour obtenir une cohérence spatiale avec la résolution de l'imagerie satellitaire, cinq pré-traitements (Figure 8.2) ont été effectué sur la donnée de référence. Tout d'abord, un rééchantillonnage (1) de la donnée de référence pour regrouper les classes en 14 classes, puis une suppression (2) des plus petits agrégats de pixels en utilisant une méthode d'analyse en composante connexe. Les *trous* issus de ce regroupement (3) ont été comblé en ap-

pliquant une méthode de plus proche voisin (PPR) puis une morphologie mathématique de fermeture (4) a permis de lisser la donnée finale. Enfin, les routes (5) ont été ajouté en fin de traitement. Les patchs de 256 pixels de côté, nécessaires pour des applications de segmentation sémantique mais aussi de classification, ont été découpé en fin de traitement sur la donnée de référence retravaillée et sur les images Sentinel-1&2. Chaque patch est accompagné d'un fichier *GeoJSON* contenant la liste des patchs Sentinel-1 et la liste des patchs Sentinel-2 associés ainsi que les classes présentes dans le patch. Au total, Multi-SenGE contient 72 033 patchs multitemporels Sentinel-2 et 1 012 227 patchs multitemporels Sentinel-1.
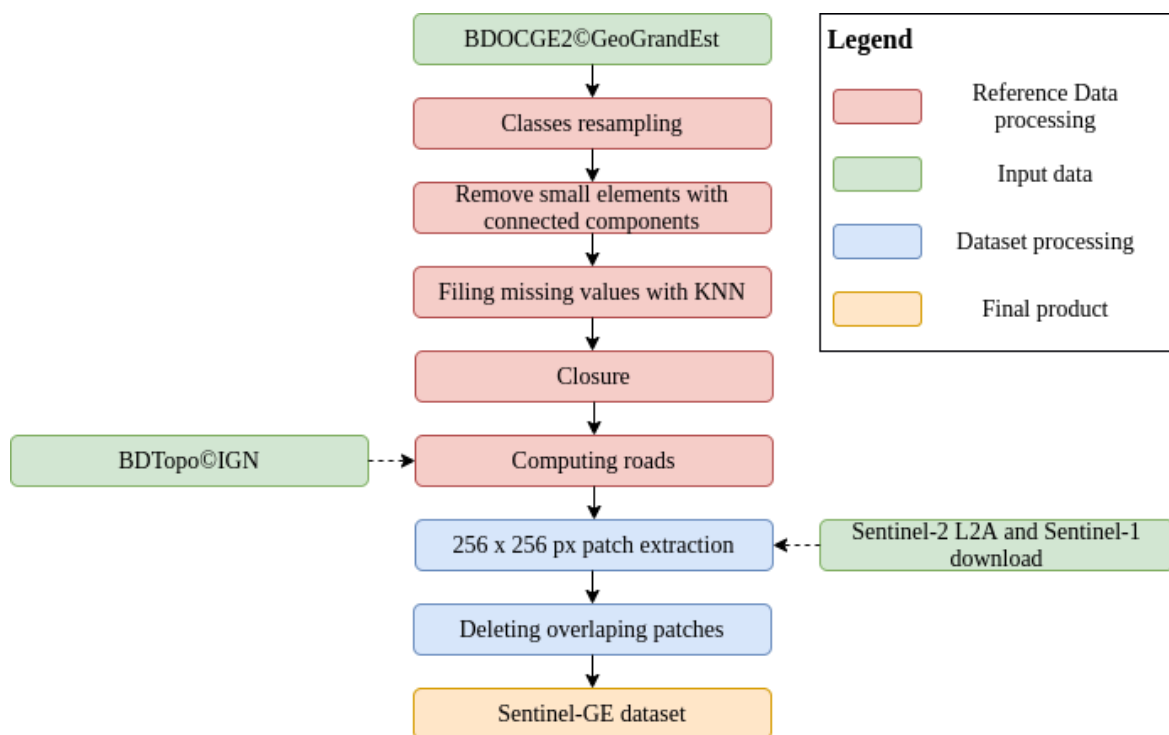

FIGURE 8.2 – *Méthodolgie pour la production du jeu de données MultiSenGE.*

Les premiers résultats ont été effectués sur de l'imagerie Sentinel-2 mono-date et uniquement sur les tissus urbains sont encourageants et ont montré l'intérêt de proposer des classifications de l'occupation des sols en incluant 4 à 5 classes de tissus urbains. Deux méthodes ont été testé en prenant comme réseau de neurones convolutionnel un U-Net [Ronneberger et al., 2015]. La première méthode, U-Net-IRRG, prend en entrée les bandes du Rouge, du Vert et du Proche Infra-Rouge. La seconde, U-Net-Index, prend en entrée les bandes du Rouge, Vert, Proche Infra-Rouge et trois indices exogènes, le NDVI, le NDBI et l'entropie calculée sur le NDVI (eNDVI). U-Net-IRRG a également été préentrainé sur ImageNet [Deng et al., 2009]. Les résultats quantitatifs et qualitatifs (Figure 8.3 et 8.3) ont

montré une amélioration pour la méthode U-Net-IRRG avec un $F1_{Score}$ global de 0.7364 et des tissus urbains plus homogènes que la méthode U-Net-Index.
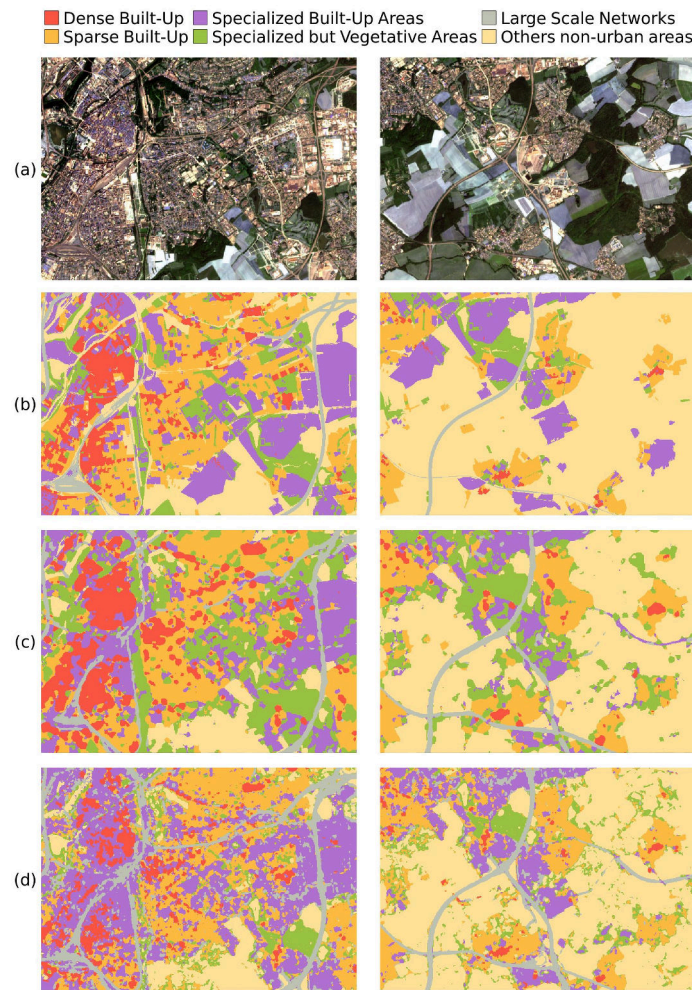

FIGURE 8.3 – *Analyse qualitative des résultats sur la zone de test de Metz. (a) est l'image Sentinel-2 en RVB, (b) est la donnée de référence en format raster, (c) classification pour U-Net-IRRG and (d) classification pour U-Net-Index.*

Weighted $F1_{Score}$	
U-Net-IRRG	U-Net-Index
0.7364	0.7214

TABLE 8.1 – *$F1_{Score}$ pondéré pour l'évaluation quantitative sur la zone de test de Metz.*

8.3 Classification multi-classes et mono-temporelle optique des tissus urbains sur la région Grand-Est

Les approches de *deep learning* ont été largement utilisé depuis plusieurs années pour la classification de l'occupation des sols et plus particulièrement pour les tissus urbains. Les méthodes de segmentation sémantique figurent parmi les plus performantes à l'heure actuelle et prennent en compte le contexte spatiale des images (photographies mais également images satellites), important dans la classification des tissus urbains ou chaque classe de tissu possède une organisation spatiale propre. Les méthodes d'encoder/decoder se révèlent être les plus performantes en apportant une délimitation fine des différents objets dans l'image [Chhor et Aramburu, 2017]. U-Net [Ronneberger et al., 2015] et SegNet [Badrinarayanan et al., 2017] sont les deux réseaux les plus performants et ont été spécifiquement développé pour de la segmentation sémantique d'images. Ces réseaux ont déjà été appliqué par le passé sur les images Sentinel-2 pour la classification des tissus urbains [El Mendili et al., 2020] et ont montré l'intérêt de classer les tissus urbains en plusieurs classes. Les indices spectraux et texturaux ont en revanche peu été intégré en tant que donnée exogènes dans des réseaux de neurones. En effet, l'approche par deep learning devrait faciliter cette extraction de caractéristiques (*Feature Extraction* ou *FX*) mais des travaux ont montré l'intérêt de l'ajout d'indice exogènes en amont ce qui améliore la convergence des réseaux [Campos-Taberner et al., 2020].

Dans ce contexte, notre étude s'est concentré sur l'évaluation de la fusion de caractéristiques entre deux réseaux convolutionels et l'ajout de caractéristiques exogènes dérivées de l'imagerie satellitaire Sentinel-2. Cette partie de thèse cherche à montrer que l'ajout d'indices exogènes en entrée de réseaux de neurones améliorent les résultats de classification des tissus urbains.

A la date de cette partie de thèse, la donnée de référence OCSGE2-GEOGRANDEST n'était pas disponible sur toute la région Grand-Est. Uniquement cinq départements nous ont été fourni par la région Grand-Est, le Bash-Rhin, les Vosges, la Moselle, la Meurthe-et-Moselle et le Haut-Rhin. Ainsi, deux tuiles Sentinel-2 ont été conservé couvrant (totalement ou en partie) ces cinq départements. Trois villes tests ont été sélectionné, Strasbourg, Metz et Saint-Avold en fonction d'un gradient d'urbanisation (Figure 8.4).
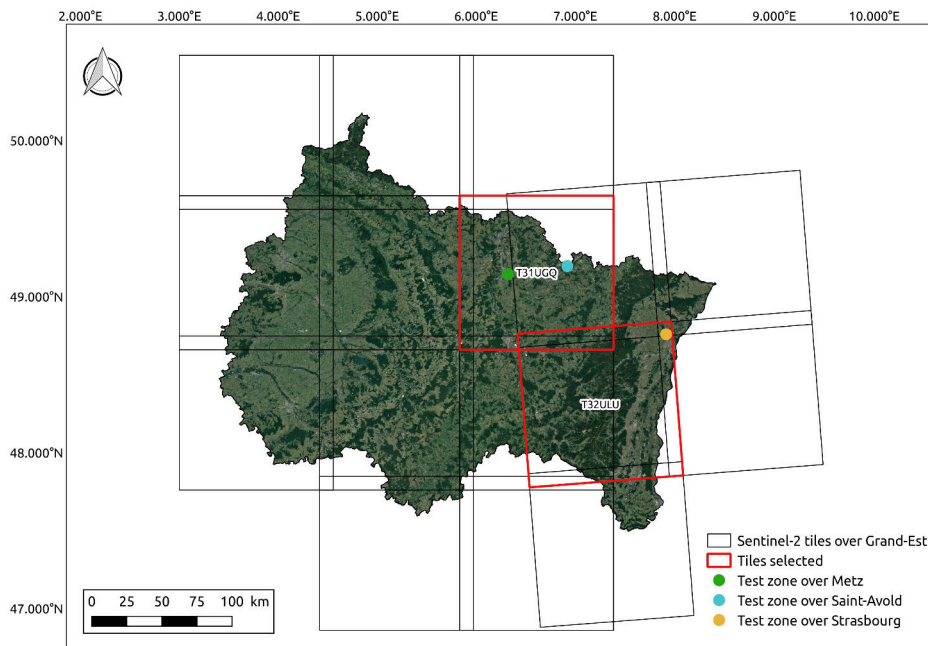
FIGURE 8.4 – *Région Grand-Est, tuiles Sentinel-2 et villes tests.*

Deux méthodes de fusion de caractéristiques entre réseaux ont été appliqué. La première consiste à fusionner les caractéristiques provenant de deux réseaux au niveau de l'encoder (U-Net-Encoder) et la seconde méthode au niveau du decoder (U-Net-Decoder). Ces deux réseaux prennent en entrée d'une part les bandes du Rouge, du Vert et du Proche Infra-rouge et d'autres part trois indices éxogènes, le NDVI, le NDBI et l'entropie calculée sur le NDVI (Figures 8.5 et 8.6). Pour comparer ces deux méthodes, les deux méthodes, développées dans la section précédentes (U-Net-IRRG et U-Net-Index), ont été également testé.
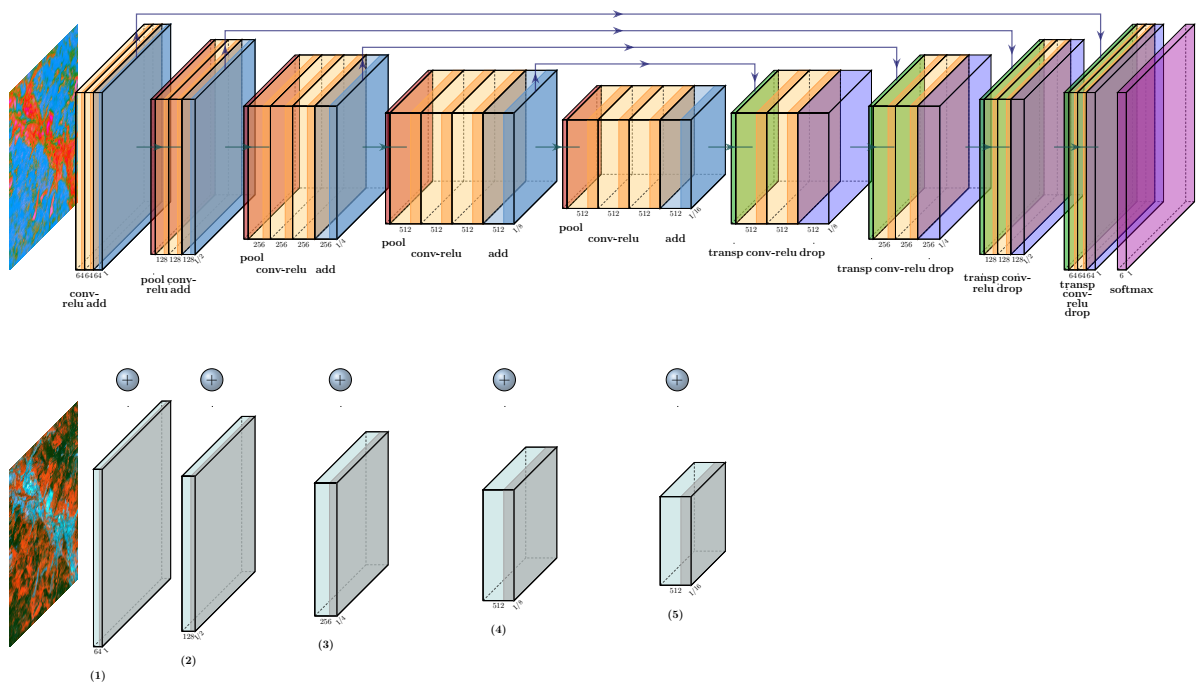
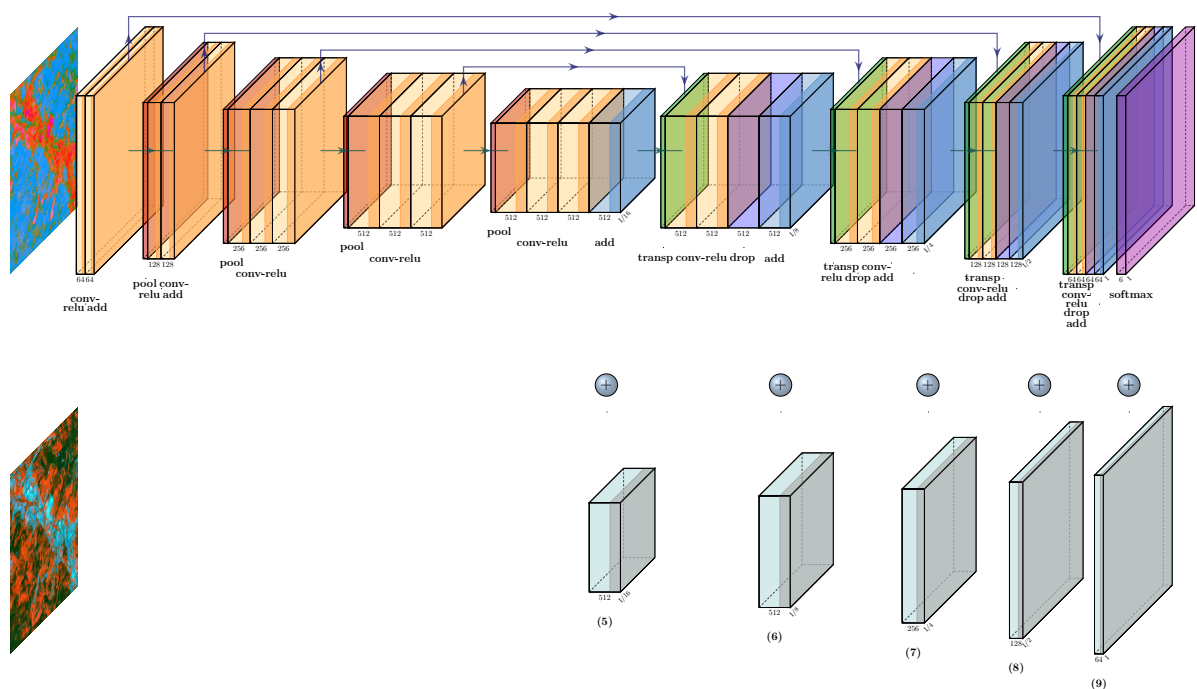FIGURE 8.5 – *Architecture U-Net-Encoder avec une fusion des caractéristiques au niveau de l'encoder des deux modèles.*

FIGURE 8.6 – *Architecture U-Net-Decoder avec une fusion des caractéristiques au niveau du decoder des deux modèles.*

Les résultats quantitatifs ont montré une meilleure classification des tissus urbains pour les deux méthodes de fusion en fonction d'un gradient d'urbanisation. U-Net-Encoder apporte les meilleurs résultats sur une ville dense telle que Strasbourg avec un $F1_{Score}$

pondéré de 0.5990 alors que U-Net-Decoder est meilleur pour les deux villes moins denses, Metz et Saint-Avold avec respectivement 0.7488 et 0.7883 de $F1_{Score}$ pondéré (Tableau 8.2).

U-Net-IRRG			U-Net-Index		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5133	0.7364	0.7716	0.5005	0.7214	0.7248
U-Net-Encoder			U-Net-Decoder		
Strasbourg	Metz	St-Avold	Strasbourg	Metz	St-Avold
0.5990	0.7479	0.7834	0.5894	**0.7488**	**0.7883**

TABLE 8.2 – *$F1_{Score}$ pondéré pour chaque méthode de la zone d'étude.*

Par rapport à l'ensemble de données de référence, l'analyse qualitative montre que les méthodes U-Net-Encoder (sous-figure e) et U-Net-Decoder (sous-figure f) permettent une meilleure détection des réseaux à grande échelle (classe 5) que les données de référence (figure 8.7 b).détection des réseaux à grande échelle (classe 5) que les données de référence (figure 8.7 b). En effet, certaines routes et voies ferrées sont classées alors qu'elles n'apparaissent pas sur l'image (Figure 8.7 b). Les Specialized Built-Up Areas (classe 3) sont également bien extraites par toutes les méthodes en détectant les zones de construction en cours qui ne sont pas présentes sur l'image ni sur la donnée de référence. D'autre part, nous remarquons une surestimation des zones de construction en cours par rapport aux données de référence. En revanche, nous remarquons une surestimation des zones Specialized But Vegetative Areas (classe 4) qui comprend certaines zones de cultures ou de forêts dans les zones périurbaines. U-Net-IRRG et U-Net-Index produisent également une certaine confusion entre les différentes classes, telles que les Specialized Built-Up Areas (classe 3) et les Specialized But Vegetative Areas (classe 4) les deux méthodes de fusion donnent des résultats beaucoup plus unifiés.
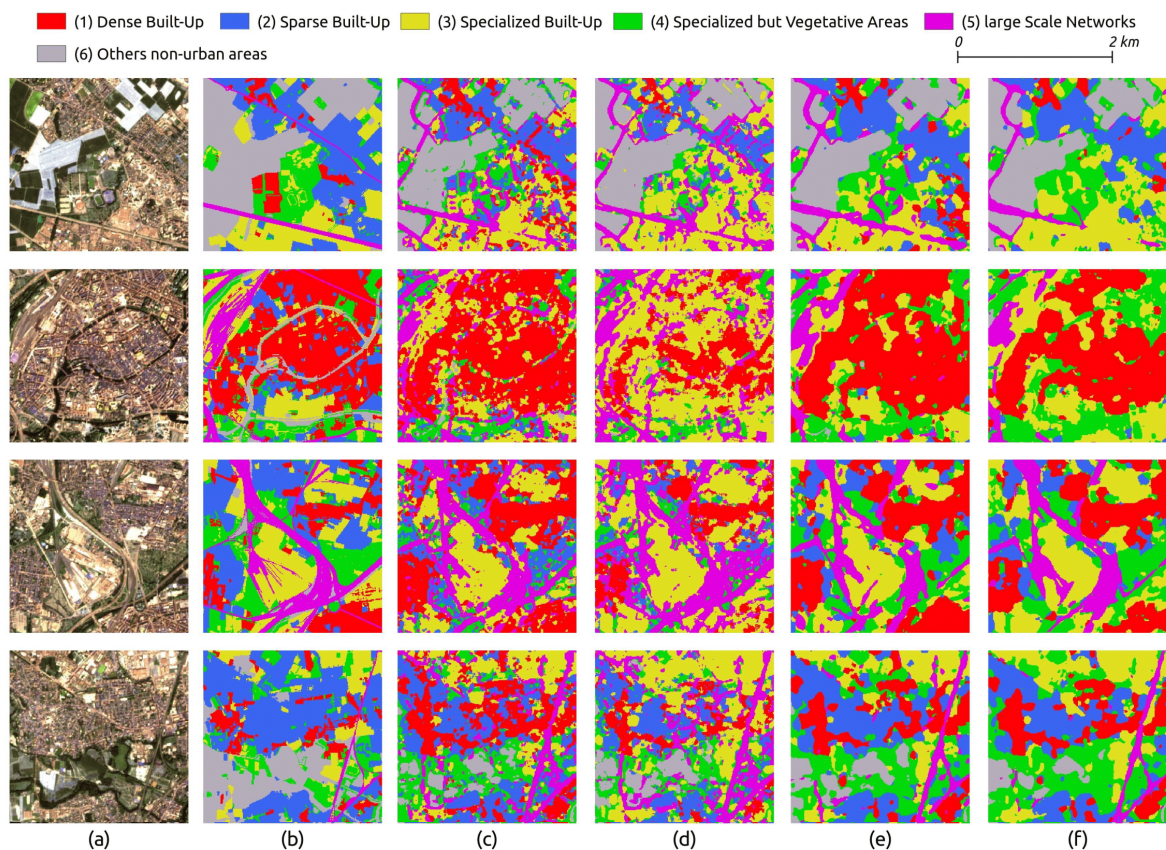
Figure 8.7 – *Exemple de résultat des quatre méthodes sur Strasbourg. (a) et (b) correspondent au subset Sentinel-2 et à la donnée de référence, (c) et (d) correspondent à U-Net-IRRG et U-Net-Index et (e) et (f) correspondent à U-Net-Encoder et U-Net-Decoder.*

L'objectif de cette recherche était de montrer l'intérêt des méthodes de segmentation sémantique pour aider les utilisateurs finaux à produire une carte pertinente et actualisée de cinq classes thématiques d'espaces urbains avec une seule image optique mono-temporelle à haute résolution spatiale (Sentinel-2). En effet, classiquement, avec une résolution spatiale élevée (10m), les espaces urbains sont uniquement cartographiés avec quatre classes qui distinguent les réseaux des zones bâties denses et clairsemées et des activités. Dans cet article, deux méthodes utilisant des techniques de fusion de caractéristiques (U-Net-Encodeur et U-Net-Décodeur) entre deux réseaux ont été développées en utilisant (i) trois bandes spectrales (vert, rouge, NIR) pour le réseau pré-entraîné et (ii) trois indices spectraux et texturaux (NDBI, NDVI et *eNDVI*) pour le réseau non pré-entraîné. Ces méthodes ont été comparées à une U-Net prenant en compte soit les bandes IRRG, soit les bandes IRRG et trois indices spectraux et texturaux. L'idée était de combiner deux types de données, chacune fournissant diverses informations pour améliorer la détection des surfaces urbaines proposées dans cinq classes différentes d'espaces urba ins adaptées aux villes du Grand-Est/France : (1) Dense Built-Up, (2) Sparse Built-Up, (3) Specialized Built-Up Areas, (4) Specialized but Vegetative Areas, (5) Large Scale Networks. Les méthodes ont

été testées en fonction d'un gradient d'urbanisation dans l'est de la France pour assurer une généralisation des résultats.

Ces résultats mettent en évidence que les deux méthodes de fusion, en particulier la méthode U-Net-Décodeur pour la plupart des espaces urbains, offrent les meilleurs résultats pour affiner la détection de la plupart des classes d'espaces urbains. La méthode U-Net-Décodeur a montré un avantage dans la délimitation des zones spécialisées mais végétales et une meilleure classification de ces zones pour les zones fortement urbanisées (centre de Strasbourg et Metz). L'analyse qualitative confirme ces premières analyses en montrant un avantage pour la méthode U-Net-Décodeur avec une meilleure segmentation des différents espaces urbains. En revanche, les résultats statistiques ont montré une classification meilleure mais proche des zones densément construites et des zones construites clairsemées pour la méthode U-Net-Encodeur. La plupart des classes d'espaces urbains sont toujours extraites avec des métriques d'évaluation pertinentes (supérieures à 0,5) et une interprétation qualitative des résultats montre des patchs d'extraction homogènes de classes. La faiblesse des résultats pour la Classe 4, qui est confondue avec d'autres classes de couverture terrestre en raison de sa complexité relative, pourrait être améliorée en utilisant des images multi-temporelles afin de tenir compte de la dynamique temporelle de la végétation.

8.4 Classification multi-modale et multi-temporelle des tissus urbains : application sur le jeu de données MultiSenGE

La section suivante a démontré l'intérêt de l'utilisation des approches de *deep learning* et plus particulièrement de segmentation sémantique [Wenger et al., 2022c] pour la classification des tissus urbains à partir d'images Sentinel-2 mono-dates. Or, l'hypothèse est émise que les variations intra-annuelles des milieux urbains et de la végétation intra-urbaine pourrait améliorer les classifications des tissus urbains. Ainsi, l'objectif de cette partie est d'explorer la combinaison des données multi-temporelles et multi-modales pour la classification des tissus urbains. Nous faisons ainsi l'hypothèse que la données (1) multi-temporelle optique et radar ainsi qu'un rééquilibrage du jeu de données améliore la classification des tissus urbains.

Le jeu de données MultiSenGE a été utilisé dans sa totalité pour cette partie de thèse. Néanmoins, il a fallu procéder à une quantification et à une extraction des patchs multi-temporels et multi-modaux. Afin de sélectionner les patchs temporels pour chaque mois, une selection multitemporelle a été effectué sur l'ensemble du jeu de données. Pour maxi-

miser le nombre de patch et la distribution spatiale des patchs, il a été retenu de séléction des patchs ayant au minimum 17 jours d'écart entre deux mois consécutifs pour les mois de juillet, août, septembre et novembre (Figure 8.8).
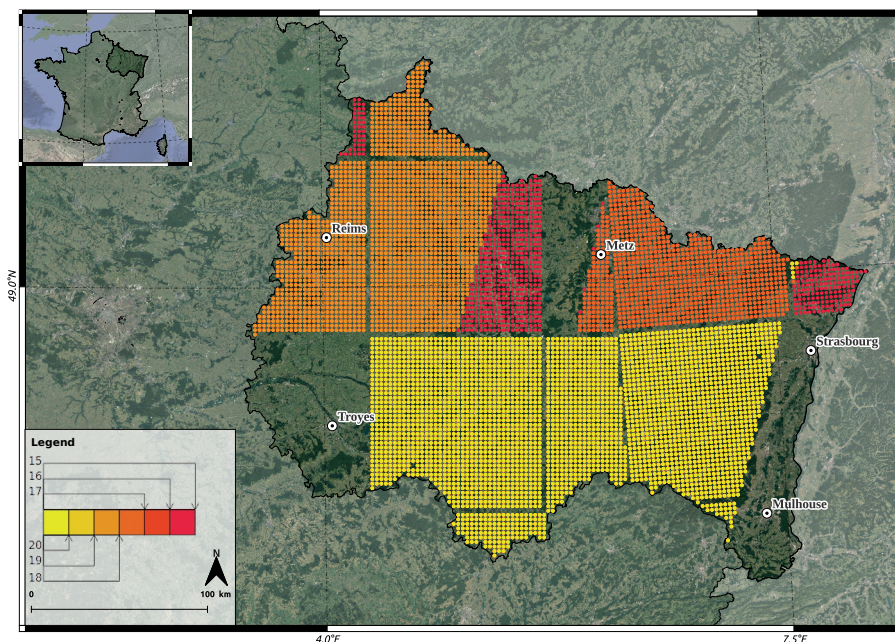

FIGURE 8.8 – *Distribution des patchs en fonction du nombre de jours entre deux mois consécutifs sur la région Grand-Est*

Le jeu de données a été découpé en jeu d'entraînement, de validation et de test en portant une attention particulière à la distribution des classes dans chaque jeu.

La donnée de référence initiale contenait 14 classes d'occupation des sols. Certaines classes ne se retrouvant pas dans tous les jeux et étant trop peu représentée à l'échelle de la région, nous avons procédé à une reclassification à 10 classes. Afin de comparer aux résultats mono-dates et afin de tester notre hypothèse, les tests ont été dupliqué pour 6 classes d'occupation des sols (5 classes urbaines et 1 classe naturelle).

Un modèle de *deep learning* multi-temporel et multi-modal a été développé. Il comprend trois branche, deux branches avec un extracteur de caractéristiques spatio-temporelles (module *ConvLSTM)*) prenant d'une part l'imagerie Sentinel-1 et d'autre part l'imagerie Sentinel-2 et un extracteur de caractéristiques spatio-spectrales (module *Inception*) prenant la concaténation de la première date Sentinel-1 et Sentinel-1 (Figure 8.9).
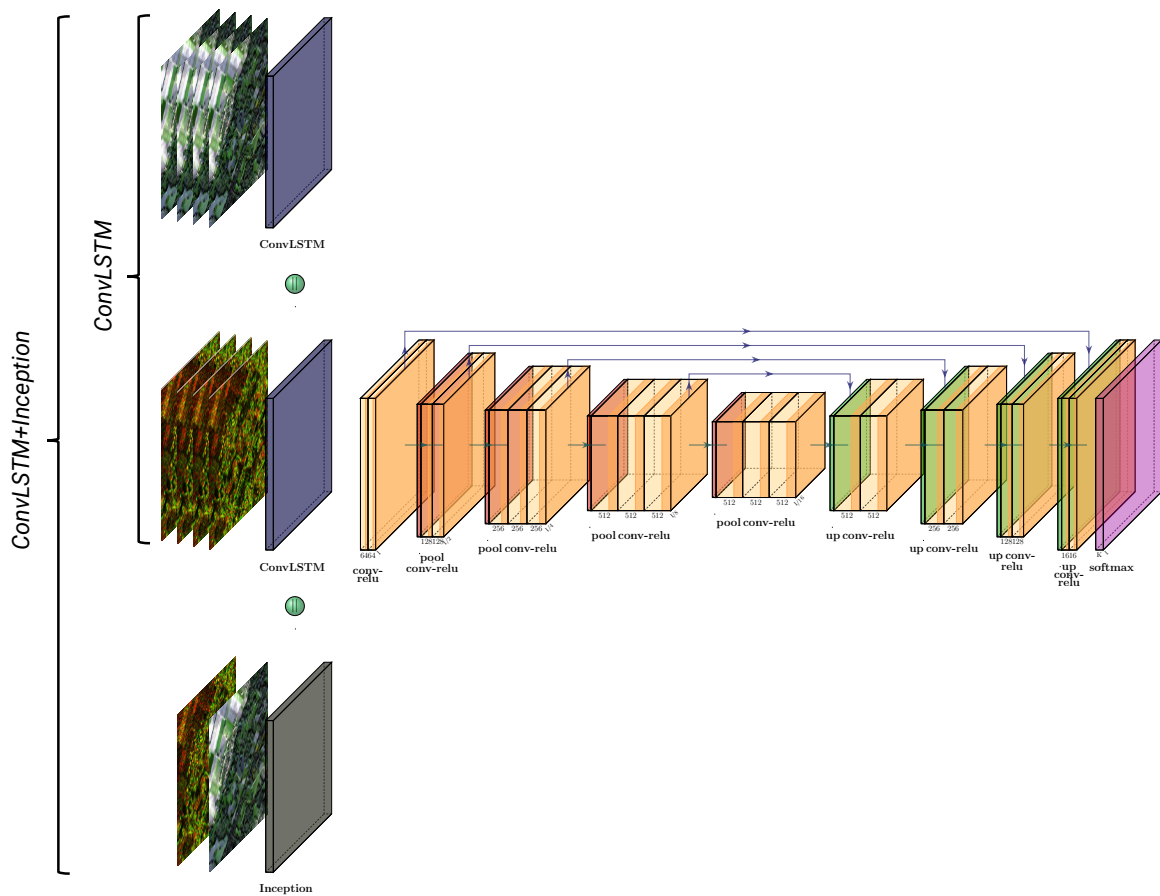
FIGURE 8.9 – *Méthode multi-temporelle et multi-modale ConvLSTM+Inception*

Les résultats à 6 classes ont montré qu'aucune méthode n'obtenait les meilleurs résultats pour la majorité des classes, même si le meilleur score Kappa global s'est retrouvé pour la méthode ConvLSTM-S2 avec 0,4223. En revanche à 10 classes, la méthode ConvLTSM+Inception-S1S2 apporte les meilleurs résultats en justifiant les choix de fusion des caractéristiques spatio-temporelles et spatio-spectrales Sentinel-1&2. En comparant les méthodes à 6 classes et la meilleur méthodes (ConvLSTM+Inception-S1S2) à 10 classes, on remarque qu'un rééquilibrage du jeu de données améliore les résultats de classification des tissus urbains (Figure 8.10).
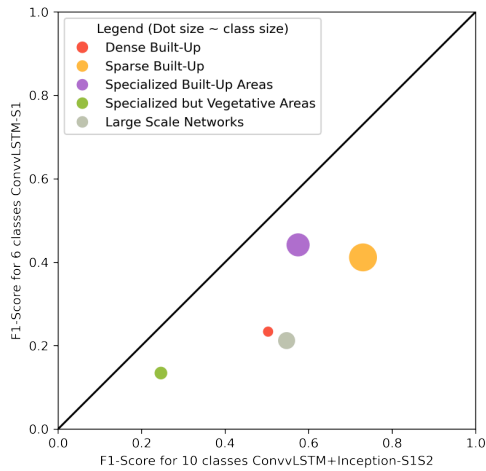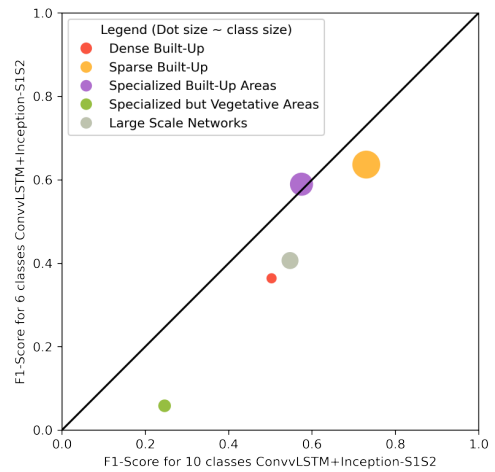
(a) 10 classes ConvLSTM+Inception-S1S2 vs 6 classes ConvLSTM-S1
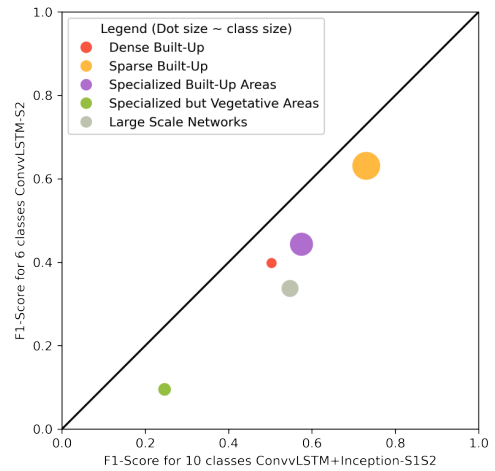
(b) 10 classes ConvLSTM+Inception-S1S2 vs 6 classes ConvLSTM-S2
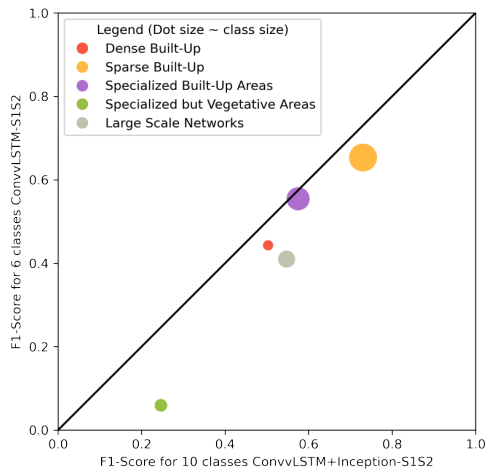
(c) 10 classes ConvLSTM+Inception-S1S2 vs 6 classes ConvLSTM-S1S2

(d) 10 classes ConvLSTM+Inception-S1S2 vs 6 classes ConvLSTM+Inception-S1S2

FIGURE 8.10 – *Graphique comparant les résultats de classification entre les méthodes à 6 classes et la meilleur méthodes à 10 classes*

En conclusion, l'imagerie multi-temporelle et muli-modale Sentinel-1&2 couplée à un rééquilibrage du jeu de données (discriminer une classe sur-représentée) permet une amélioration des classification des tissus urbains. Ces modèles permettent la production de cartographies pouvant être incluses dans des modèles climatiques pour caractériser les *Local Climate Zone* (LCZ).

8.5 INFÉRENCE TEMPORELLE ET SPATIALE

Les recherches actuelles en *deep learning* et en télédétection se contentent souvent d'entraîner et d'appliquer des modèles sur un jeu de données défini (par exemple : MultiSenGE).

Or, dans un objectif de cartographie à grande échelle, il pourrait être intéressant d'évaluer la capacité de généralisation de ces réseaux entraînés sur un jeu de données existant et appliqué sur des territoires situés à plusieurs centaines de kilomètres de la région d'entraînement. Dans cette dernière partie de thèse, l'inférence spatiale et l'inférence temporelle du modèle ConvLSTM+Inception-S1S2 développé dans la partie précédente ont été exploré pour les espaces urbains (5 classes).

L'inférence spatiale consiste à explorer les limites d'un modèle entraîné sur une région spécifique, dans notre cas le Grand-Est, sur d'autres villes en France et en Europe en fonction d'un critère d'éloignement. Cinq villes en France et trois en Europe ont donc été choisi : Toulouse, Dijon, Orléans, Lille et Rennes ainsi que Séville (Espagne), Oran (Algérie) et Brême (Allemagne). L'évaluation des classifications pour les villes Française ont été faite par digitalisation manuelle de cinq zones de 4 km^2 par ville en allant du centre vers la périphérie poiur couvrir au mieux les différents morphologies urbaines (Annexes A.14, A.15, A.16, A.17, A.18, A.19). Cette technique d'inférence spatiale permettrait la production de cartographie à large échelle à moindre coût, tant humain que financier.

L'inférence temporelle consiste à explorer la détection de changement à deux dates sur la même région d'entraînement entre deux classifications. Ainsi, le modèle ConvLSTM+Inception-S1S2 a été appliqué sur le Grand-Est pour 2018 et 2022 pour produire deux cartes d'occupation des sols. L'objectif est de détecter et d'identifier les changements brusques des tissus urbains (espaces naturels vers espaces urbains) pour proposer un système d'alerte.

Les résultats de l'inférence spatiale ont montré que plus nous nous éloignons de la région d'entraînement moins bon seront les résultats (Table 8.3 et Figure 8.11). L'hypothèse est ainsi émise que plus nous nous éloignons de la région d'entraînement, plus la morphologie des villes diffère des villes de la région d'entraînement.

Table 8.3 – *$F1_{Score}$ pondéré pour chaque ville.*

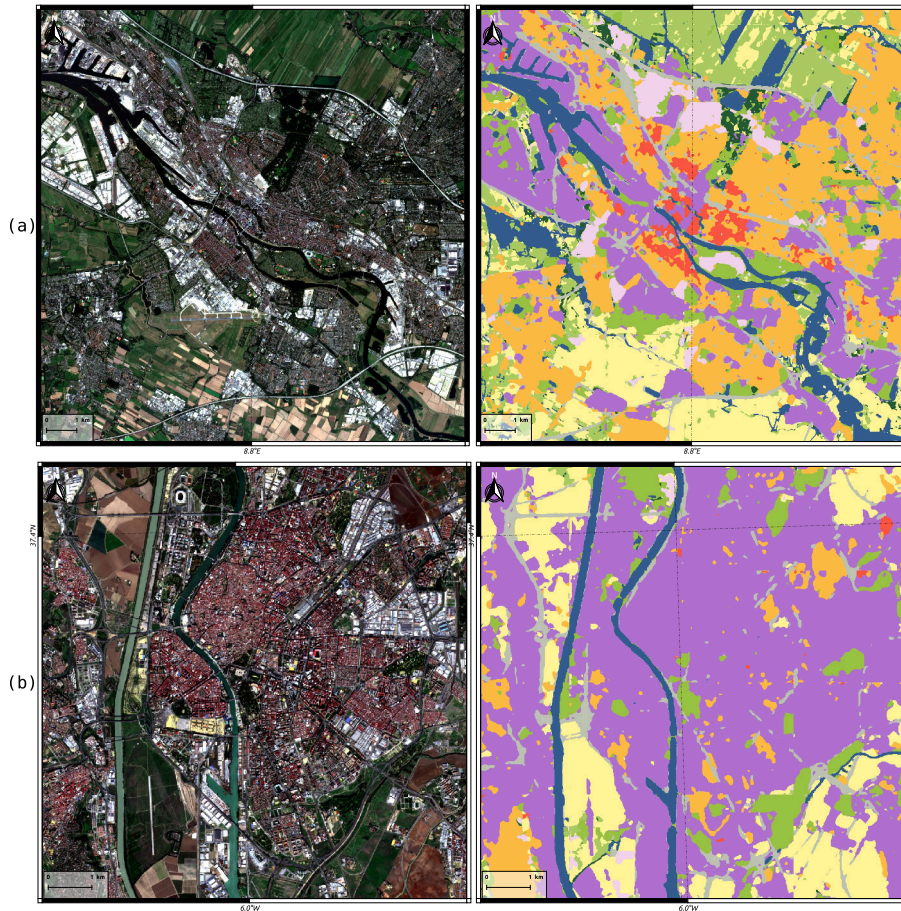| Villes	Score
Toulouse	0.7029
Dijon	**0.8087**
Orléans	0.7742
Lille	0.6866
Rennes	0.7187

FIGURE 8.11 – *Résultats de classification pour Brème (a) et Séville (b).*

Les résultats d'inférence temporelle ont permis d'isoler de nombreux changements sur la région Grand-Est (Figure 8.12). En revanche, cette méthode ne permet pas de quantifier le changement, qui est soumis à de trop nombreux faux positifs dans les résultats de classification. Néanmoins, ce système d'alerte peut permettre aux acteurs des territoires de mettre à jour plus facilement les cartographies manuelles des milieux urbains.
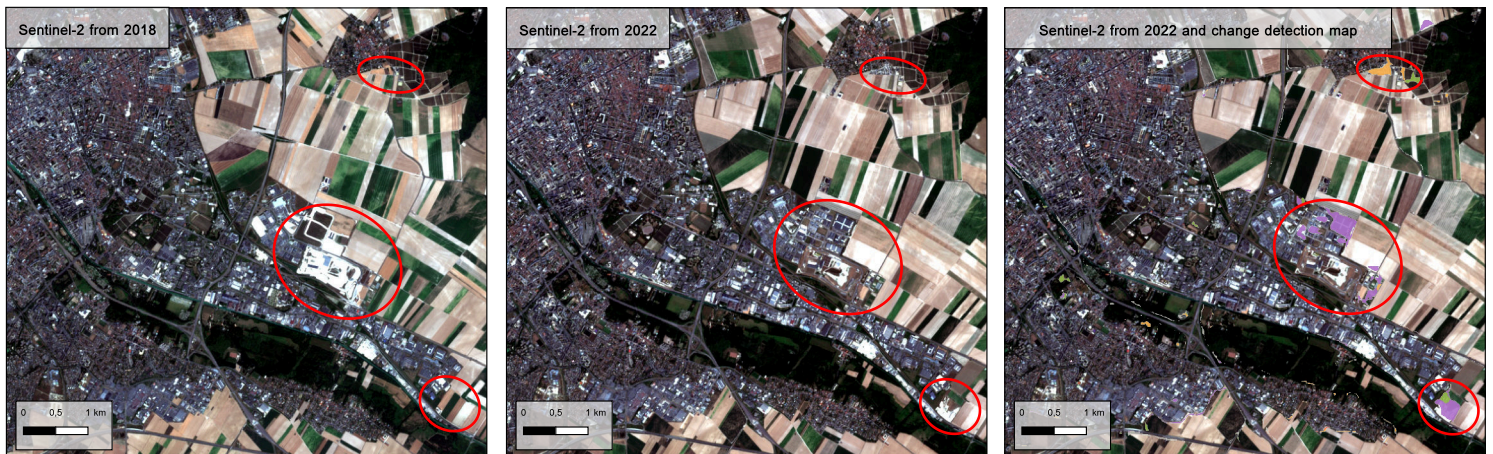

FIGURE 8.12 – *Détection de changements proche de la ville de Reims.*

8.6 Conclusion

L'objectif principal de cette thèse était l'évaluation du potentiel de l'imagerie multi-temporelle optique et radar à haute résolution spatiale pour la cartographie de l'occupation des sols (en particulier des tissus urbains) pour analyser leurs dynamiques à grande échelle afin d'aider les acteurs des territoires à obtenir une donnée à jour pour leurs applications.

Les principales contribution de cette thèse peuvent être résumées en quatre points :

- La construction d'un jeu de données multi-temporel et multi-modal innovant qui a été proposé à la communauté scientifique.

- La formalisation de l'utilisation de *deep learning* et de la segmentation sémantique pour la classification des tissus urbains.

- L'intégration de données multi-temporelles et multi-modales qui ont permis d'améliorer les résultats de classification mono-dates des tissus urbains.

- La généralisation des approches avec l'inférence temporelle et spatiale pour la détection de changements brusques et la cartographie à grande échelle.

Plusieurs perspectives sont ainsi soulevé à l'issue de cette thèse. La première consiste en l'amélioration du jeu de données MultiSenGE en ajoutant une données à très haute résolution spatiale mono-date (type Pléiades ou SPOT) pour améliorer les résultats de classification. La seconde consiste à étoffer la donnée de référence, notamment pour les surfaces naturelles afin de distinguer plusieurs types de cultures. Enfin, la dernière perspective consisterait à proposer une cartographie à grande échelle en combinant les produits 'urbains' de cette thèse avec les classes naturelles du produit OSO.

Appendices

<div style="text-align: right">A</div>

Contents

A.1 Chapter 1: Typology for OSO, OCSGE2, Urban Atlas and Corine Land Cover

(a)

- Urbain Dense
- Urbain Diffus
- Zones industrielles et commerciales
- Routes
- Colza
- Céréales à paille
- Protéagineux
- Soja
- Tournesol
- Maïs
- Riz
- Tubercules/Racines
- Prairies
- Vergers
- Vignes
- Forêts de feuillus
- Forêts de conifères
- Pelouse
- Landes
- Surfaces minérales
- Plages et Dunes
- Glaciers et neiges éternelles
- Eau

(b)

- Bati continu dense
- Bati continu aere
- Bati collectif
- Bati mixte
- Bati individuel dense
- Bati individuel lache
- Bati isolé en zone agricole ou naturelle
- Espaces libres en milieu urbain
- Emprises scolaires et universitaires
- Emprises hospitalieres
- Equipements sportifs et de loisirs campings
- Cimetieres
- Autres equipements collectifs
- Equipements eau, energies, T.I.C. et dechets
- Emprises d'activites a dominante commerciale
- Emprises d'activites a dominante mixte ou tertiaire
- Anciennes emprises d'activite
- Emprises d'activites a dominante industrielle
- Emprises militaires
- Exploitations agricoles
- Zones d'extraction
- Emprise reseau ferre
- Emprise reseau routier
- Espaces associes aux reseaux routiers et ferres
- Emprises aeroportuaires
- Emprises portuaires
- Espaces verts urbains
- Espaces en transition
- Places
- Cultures annuelles et pluri-annuelles
- Cultures specifiques
- Surfaces enherbees, friches et delaissés agricoles
- Vignes
- Vergers traditionnels
- Vergers intensifs
- Pepinieres
- Bosquets et haies
- Forêts de feuillus
- Forêts de coniferes
- Forêts mixtes
- Peuplerales et sapinieres
- Coupes a blanc et jeunes plantations
- Pelouses et pâturages de montagne
- Formations pre-forestieres
- Surfaces enherbees semi-naturelles
- Plages et sables
- Plages et sables
- Roches nues
- Zones de sinistre (incendie, tempete)
- Ripisylves et rivulaires
- Autres milieux humides
- Cours d'eau et canaux
- Plans d'eau
- Bassins artificiels
- Permeable
- impermeable bati
- impermeable non bati
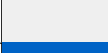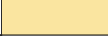
(c)

- 11100: Continuous Urban fabric (S.L. > 80%)
- 11210: Discontinuous Dense Urban Fabric (S.L.: 50% - 80%)
- 11220: Discontinuous Medium Density Urban Fabric (S.L.: 30% - 50%)
- 11230: Discontinuous Low Density Urban Fabric (S.L.: 10% - 30%)
- 11240: Discontinuous very low density urban fabric (S.L.: < 10%)
- 11300: Isolated Structures
- 12100: Industrial, commercial, public, military and private units
- 12210: Fast transit roads and associated land
- 12220: Other roads and associated land
- 12230: Railways and associated land
- 12300: Port areas
- 12400: Airports
- 13100: Mineral extraction and dump sites
- 13300: Construction sites
- 13400: Land without current use
- 14100: Green urban areas
- 14200: Sports and leisure facilities
- 21000: Arable land (annual crops)
- 22000: Permanent crops
- 23000: Pastures
- 24000: Complex and mixed cultivation patterns
- 25000: Orchads
- 31000: Forests
- 32000: Herbaceous vegetation associations
- 33000: Open spaces with little or no vegetations
- 40000: Wetlands
- 50000: Water

(d)
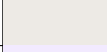
- 111 - Tissu urbain continu
- 112 - Tissu urbain discontinu
- 121 - Zones industrielles ou commerciales et installations publiques
- 122 - Réseaux routier et ferroviaire et espaces associés
- 123 - Zones portuaires
- 124 - Aéroports
- 131 - Extraction de matériaux
- 132 - Décharges
- 133 - Chantiers
- 141 - Espaces verts urbains
- 142 - Equipements sportifs et de loisirs
- 211 - Terres arables hors périmètres d'irrigation
- 212 - Périmètres irrigués en permanence
- 213 - Rizières
- 221 - Vignobles
- 222 - Vergers et petits fruits
- 223 - Oliveraies
- 231 - Prairies et autres surfaces toujours en herbe à usage agricole
- 241 - Cultures annuelles associées à des cultures permanentes
- 242 - Systèmes culturaux et parcellaires complexes
- 243 - Surfaces essentiellement agricoles, interrompues par des espaces naturels importants
- 244 - Territoires agroforestiers
- 311 - Forêts de feuillus
- 312 - Forêts de conifères
- 313 - Forêts mélangées
- 321 - Pelouses et pâturages naturels
- 322 - Landes et broussailles
- 323 - Végétation sclérophylle
- 324 - Forêt et végétation arbustive en mutation
- 331 - Plages, dunes et sable
- 332 - Roches nues
- 333 - Végétation clairsemée
- 334 - Zones incendiées
- 335 - Glaciers et neiges éternelles
- 411 - Marais intérieurs
- 412 - Tourbières
- 421 - Marais maritimes
- 422 - Marais salants
- 423 - Zones intertidales
- 511 - Cours et voies d'eau
- 512 - Plans d'eau
- 521 - Lagunes littorales
- 522 - Estuaires
- 523 - Mers et océans

A.2 Chapter 1: Typology for ESA World Cover

ESA World Cover semantic classes	LCCS Code	Color
Tree cover (10)	A12A3 // A11A1 A24A3C1(C2)-R1(R2)	
Shrubland (20)	A12A4 // A11A2	
Grassland (30)	A12A2	
Cropland (40)	A11A3(A4)(A5) // A23	
Built-up (50)	B15A1	
Bare/sparse vegetation (60)	B16A1(A2) // B15A2	
Snow and Ice (70)	B28A2(A3)	
Permanent water bodies (80)	B28A1(B1) // B27A1(B1)	
Herbaceous wetland (90)	A24A2	
Mangroves (95)	A24A3C5-R3	
Moss and lichen (100)	A12A7	

Numbers in parenthesis represent the pixel value in the raster classification. To obtain a better readability, ESA World Cover linked the LCCS code corresponding to each semantic class. More informations about this LULC product can be found following this link: https://esa-worldcover.org/en

A.3 Chapter 1: Typology for ESRI 2020 Land Cover

ESRI 2020 Land Cover semantic classes	Color
Water (1)	
Trees (2)	
Grass (3)	
Flooded Vegetation (4)	
Crops (5)	
Scrub/Shrub (6)	
Built Area (7)	
Bare Ground (8)	
Snow/Ice (9)	
Clouds (10)	

Numbers in parenthesis represent the pixel value in the raster classification. More informations about this LULC product can be found following this link: https://livingatlas.arcgis.com/landcover/

A.4 CHAPTER 1: TYPOLOGY FOR GOOGLE DYNAMIC WORLD

Google Dynamic World semantic classes	Color
Water (1)	
Trees (2)	
Grass (3)	
Flooded Vegetation (4)	
Crops (5)	
Shrub & Scrub (6)	
Built Area (7)	
Bare Ground (8)	
Snow & Ice (9)	

Numbers in parenthesis represent the pixel value in the raster classification. More informations about this LULC product can be found following this link: https://dynamicworld.app/

A.5 CHAPTER 5: BAR PLOT RESULTS OF ALL METHODS FOR THE TEST ZONE LOCATED IN THE WEST OF THE GRAND-EST REGION FOR 6 SEMANTIC CLASSES
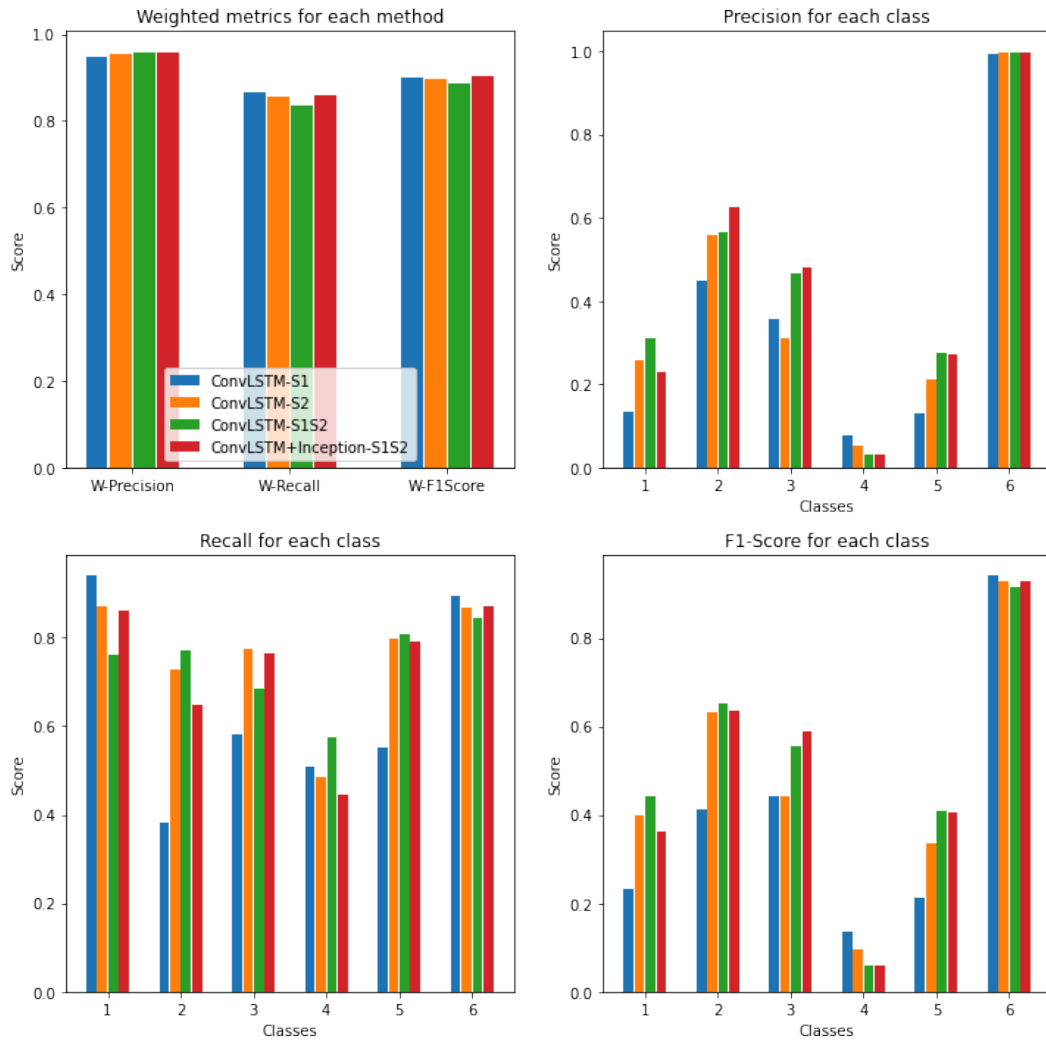
A.6 Chapter 5: Confusion matrix computed over the test dataset for every method for 6 semantic LULC classes. (a) Confusion matrix for ConvLSTM-S1. (b) Confusion matrix for ConvLSTM-S2. (c) Confusion matrix for ConvLSTM-S1S2. (d) Confusion matrix for ConvLSTM+Inception-S1S2.
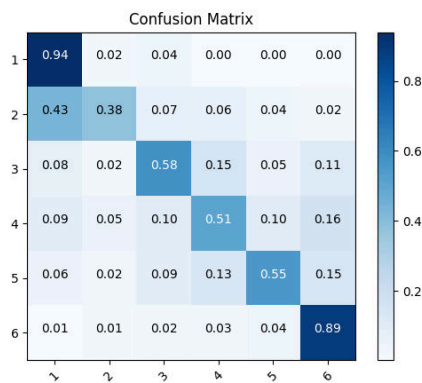
(a)
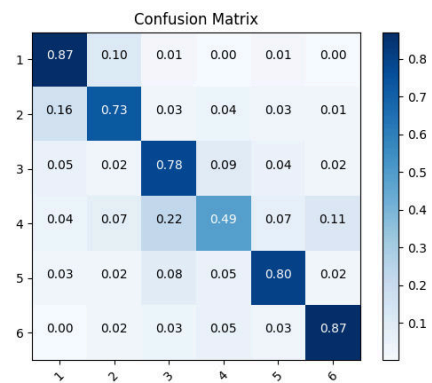
(b)
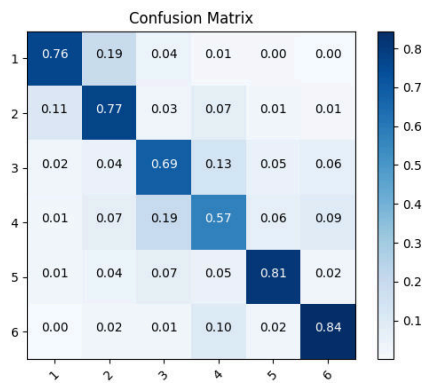
(c)
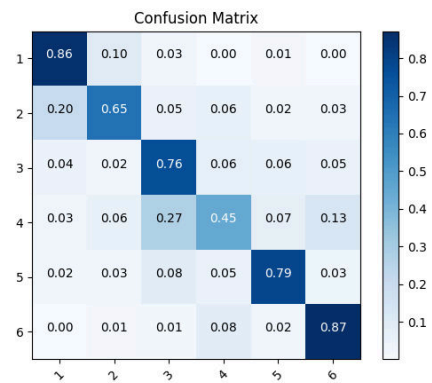
(d)

A.7 CHAPTER 5: BAR PLOT RESULTS OF ALL METHODS FOR THE TEST ZONE LOCATED IN THE WEST OF THE GRAND-EST REGION FOR 10 SEMANTIC CLASSES.
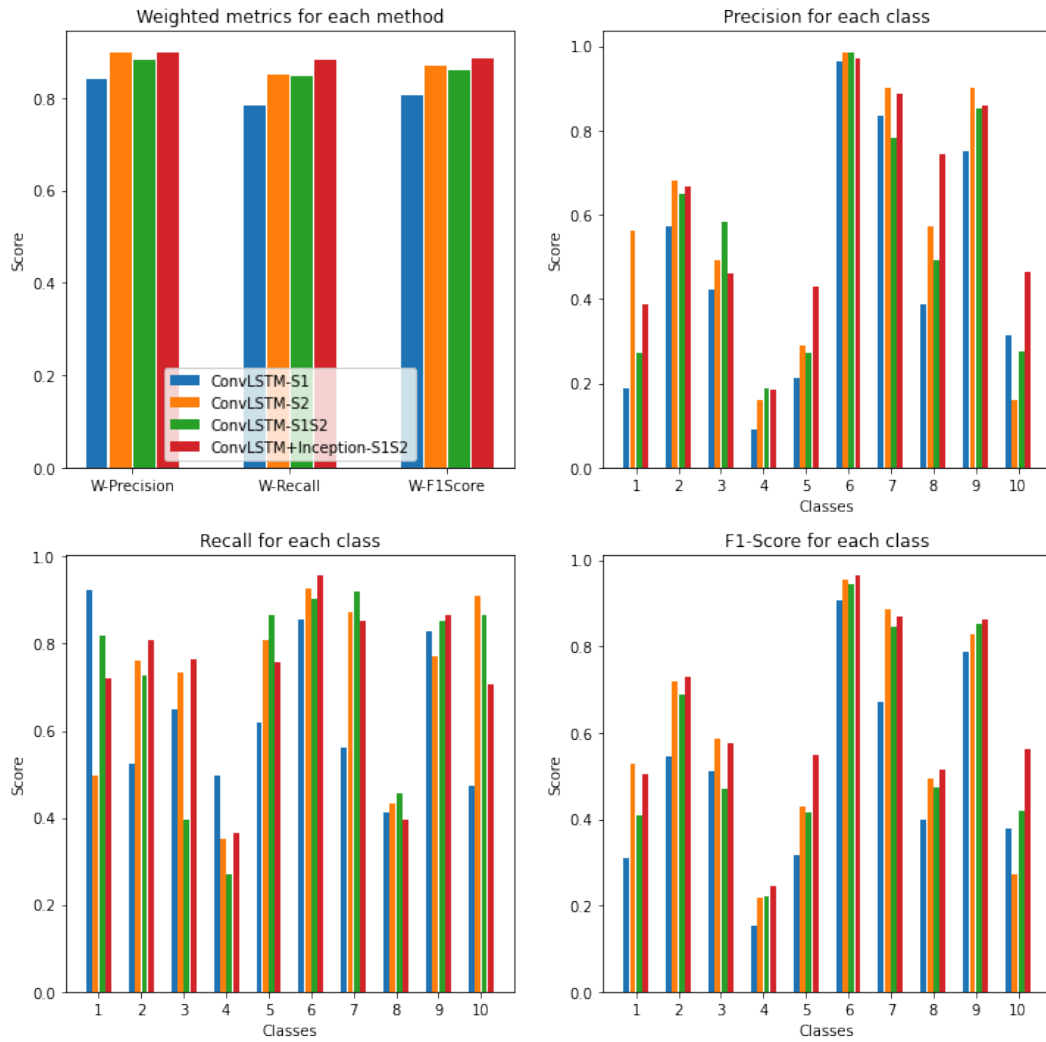
A.8 CHAPTER 5: CONFUSION MATRIX COMPUTED OVER THE TEST DATASET FOR EVERY METHOD FOR 10 SEMANTIC LULC CLASSES. (A) CONFUSION MATRIX FOR CONVLSTM-S1. (B) CONFUSION MATRIX FOR CONVLSTM-S2. (C) CONFUSION MATRIX FOR CONVLSTM-S1S2. (D) CONFUSION MATRIX FOR CONVLSTM+INCEPTION-S1S2.
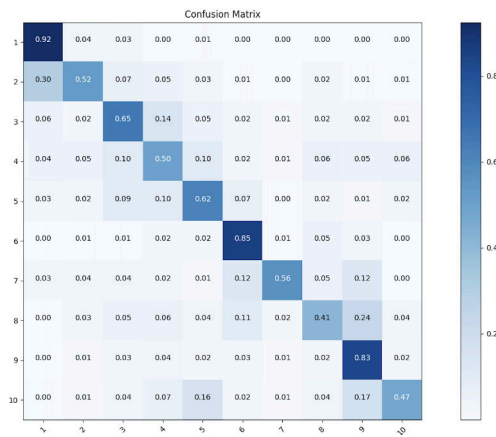
(e)
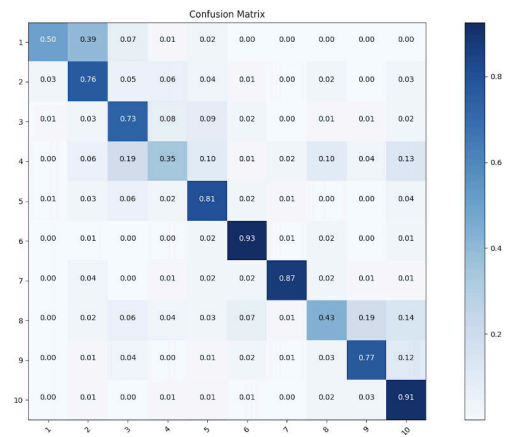
(f)

(g)
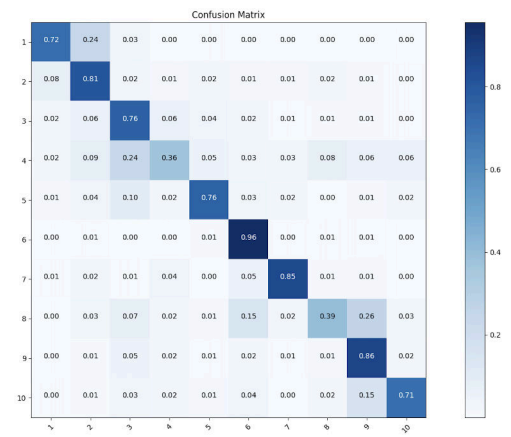
(h)

A.9 Chapter 6: Sentinel-2 date selection for temporal inference over Grand-Est region for 2018

Tile	Date 1 (∼ July)	Date 2 (∼ August)	Date 3 (∼ September)	Date 4 (∼ November)
T31TFN	27 July 2018	16 August 2018	25 September 2018	20 October 2018
T31UEP	15 July 2018	18 August 2018	28 September 2018	13 October 2018
T31UEQ	25 July 2018	04 August 2018	8 September 2018	17 November 2018
T31UER	25 July 2018	04 August 2018	18 September 2018	17 November 2018
T31UFP	27 July 2018	16 August 2018	20 September 2018	15 October 2018
T31UFQ	27 July 2018	16 August 2018	25 September 2018	20 October 2018
T31UFR	2 July 2018	16 August 2018	10 October 2018	4 November 2018
T31UGP	2 July 2018	6 August 2018	20 September 2018	15 October 2018
T31UGQ	27 July 2018	16 August 2018	25 September 2018	20 October 2018
T32TLT	29 July 2018	28 August 2018	27 September 2018	22 October 2018
T32ULU	9 July 2018	3 August 2018	27 September 2018	12 October 2018
T32ULV	24 July 2018	18 August 2018	27 September 2018	22 October 2018
T32UMU	9 July 2018	3 August 2018	17 September 2018	17 October 2018
T32UMV	9 July 2018	3 August 2018	27 September 2018	11 November 2018

The *ConvLSTM+Inception-S1S2* model was trained with patches from July, August, September and November. To respect the same selection of dates, each image from the Grand-Est region, with less than 30% cloud cover, was trained. The classification was done for all tiles.

A.10 Chapter 6: Sentinel-2 date selection for temporal inference over Grand-Est region for 2022

Tile	Date 1 (∼ April)	Date 2 (∼ May)	Date 3 (∼ June)	Date 4 (∼ August)
T31TFN	28 March 2022	17 May 2022	11 June 2022	25 August 2022
T31UEP	20 April 2022	15 May 2022	19 June 2022	13 August 2022
T31UEQ	26 March 2022	15 May 2022	4 July 2022	13 August 2022
T31UER	26 March 2022	15 May 2022	4 June 2022	13 August 2022
T31UFP	22 April 2022	17 May 2022	16 June 2022	25 August 2022
T31UFQ	17 April 2022	17 May 2022	16 June 2022	25 August 2022
T31UFR	17 April 2022	22 May 2022	16 June 2022	10 August 2022
T31UGP	17 April 2022	17 May 2022	6 July 2022	25 August 2022
T31UGQ	17 April 2022	17 May 2022	16 June 2022	25 August 2022
T32TLT	25 March 2022	14 May 2022	18 June 2022	12 August 2022
T32ULU	25 March 2022	14 May 2022	18 June 2022	12 August 2022
T32ULV	25 March 2022	14 May 2022	18 June 2022	22 August 2022
T32UMU	19 April 2022	19 May 2022	18 June 2022	12 August 2022
T32UMV	25 March 2022	14 May 2022	18 June 2022	12 August 2022

For the year 2022, the same selection of dates as the model training was not possible due to delays. Thus, we have chosen for each tile of the Grand-Est one date per month for the months of April, May, June and August with a cloud cover lower than 30%.

A.11 Chapter 6: Sentinel-2 date selection for spatial inference over France for 2020

Tile	Date 1 (∼ July)	Date 2 (∼ August)	Date 3 (∼ September)	Date 4 (∼ November)
T31TCJ	9 July 2020	8 August 2020	17 September 2020	17 October 2020
T31TFN	31 July 2020	5 August 2020	14 September 2020	28 November 2020
T31UDP	9 July 2020	29 July 2020	17 September 2020	21 November 2020
T31UES	24 June 2020	13 August 2020	17 September 2020	6 November 2020
T30UWU	30 July 2020	9 August 2020	13 September 2020	17 December 2020

A.12 Chapter 6: Performing Sankey plot for each subset when a natural pixel turns to urban between 2018 and 2022

Algorithm 0: Performing Sankey plot for each subset when a natural pixel turns to urban between 2018 and 2022

Inputs: $2018_{0,0}...2018_{x,y}$, $2022_{0,0}...2022_{x,y}$ (2018 and 2022 classification matrix)

Output: `Dic` (dictionnary quantifying changes)

Require: $shape(2018) = shape(2022)$

i, j $\leftarrow 0$

x, y $\leftarrow shape(2018)$

urban_values $\leftarrow [1, 2, 3, 4, 5]$

natural_values $\leftarrow [6, 7, 8, 9, 10]$

while i $\leq$ x **do**

  while j $\leq$ y **do**

    if 2018[i, j] $\neq$ 2022[i, j] **then**

      if 2018[i, j] $\in$ natural_values **and** 2022[i, j] $\in$ urban_values **then**

        Dic[2018[i, j], 2022[i, j] $\leftarrow$ Dic[2018[i, j], 2022[i, j] $+ 1$

      end if

    end if

    j $\leftarrow$ j $+ 1$

  end while

  i $\leftarrow$ i $+ 1$

end while

A.13 Chapter 6: Algorithm to map each pixel change from natural to urban fabric

---

Algorithm 0: Algorithm to map each pixel change from natural to urban fabric

---

Inputs: $2018_{0,0}...2018_{x,y}$, $2022_{0,0}...2022_{x,y}$ (2018 and 2022 classification matrix)

Output: `change_map` (Map with each pixel changing from natural to urban)

Require: $shape(2018) = shape(2022)$

$\texttt{i,j} \leftarrow 0$

$\texttt{x,y} \leftarrow shape(2018)$

$\texttt{urban\_values} \leftarrow [1,2,3,4,5]$

$\texttt{natural\_values} \leftarrow [6,7,8,9,10]$

while $\texttt{i} \leq \texttt{x}$ **do**

  while $\texttt{j} \leq \texttt{y}$ **do**

    if $\texttt{2018[i, j]} \neq \texttt{2022[i, j]}$ **then**

      if $\texttt{2018[i, j]} \in \texttt{natural\_values}$ **and** $\texttt{2022[i, j]} \in \texttt{urban\_values}$ **then**

        $\texttt{change\_map[i, j]} \leftarrow \texttt{2022[i, j]}$

      end if

    end if

    $\texttt{j} \leftarrow \texttt{j} + 1$

  end while

  $\texttt{i} \leftarrow \texttt{i} + 1$

end while

A.14 Chapter 6: Digitization areas and results for Orléans
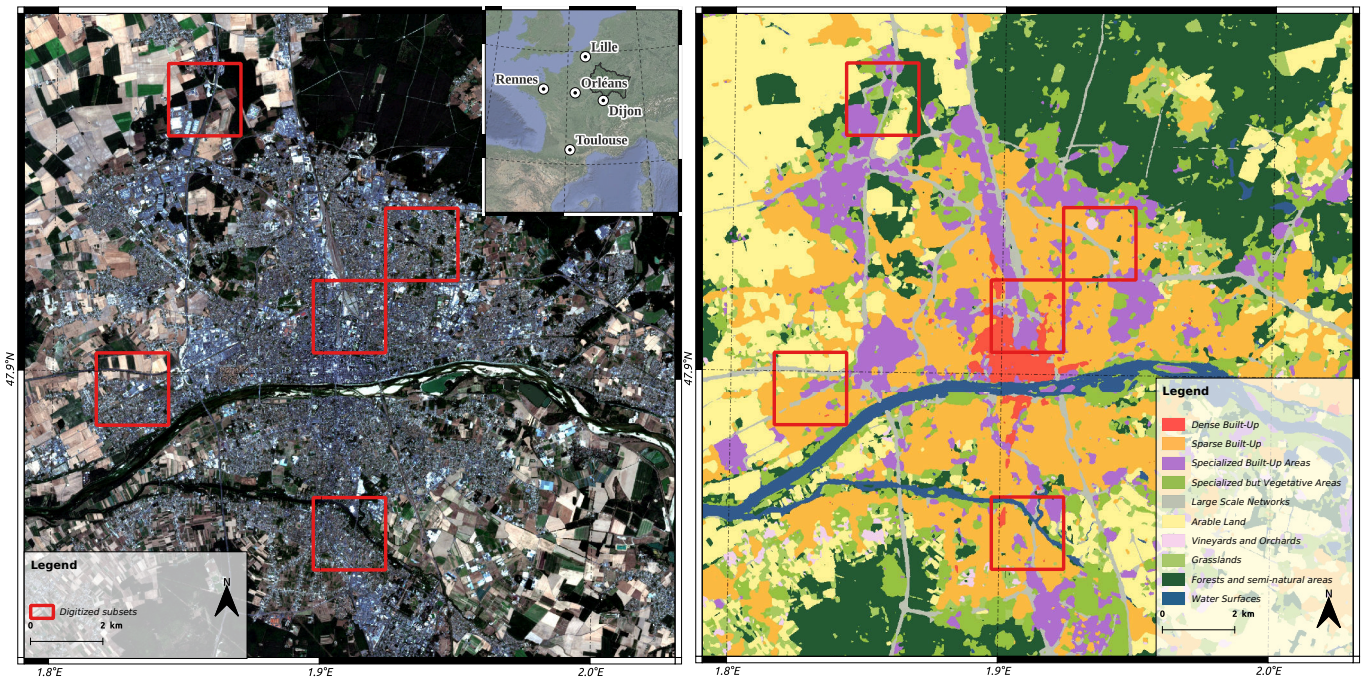
A.15 Chapter 6: Digitization areas and results for Toulouse
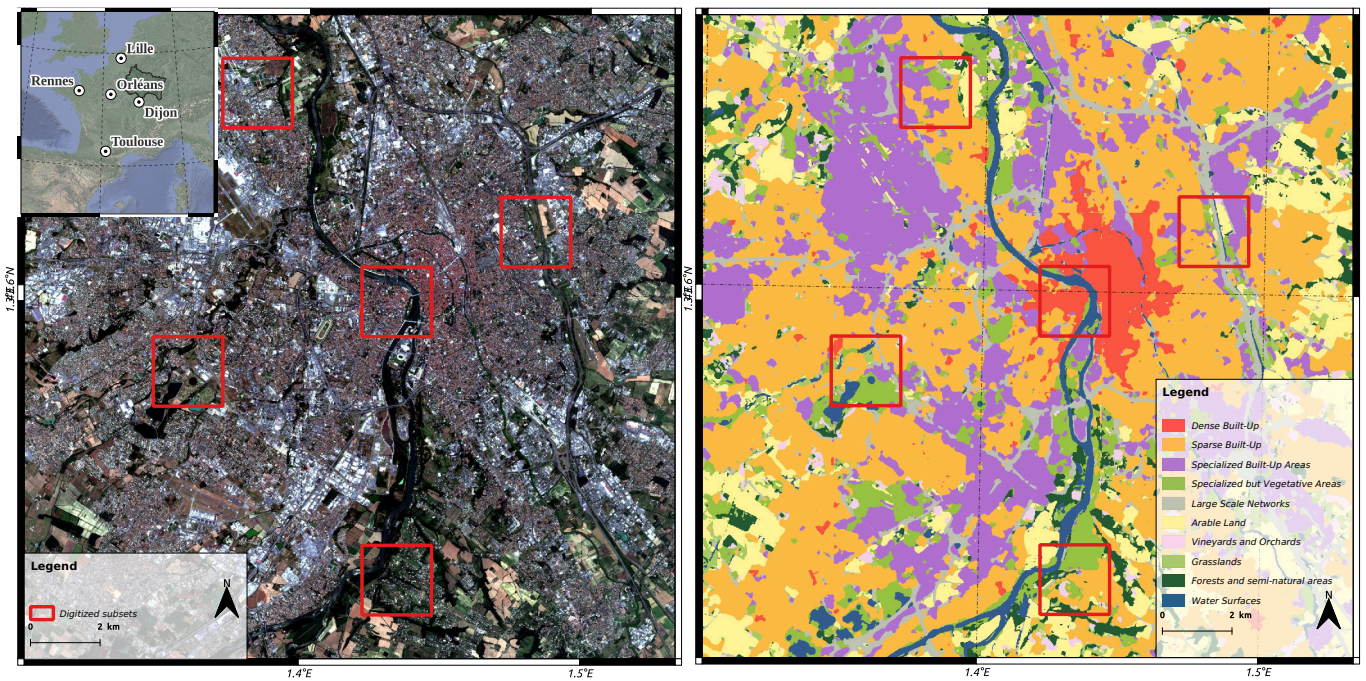
A.16 CHAPTER 6: DIGITIZATION EXAMPLE FOR TOULOUSE
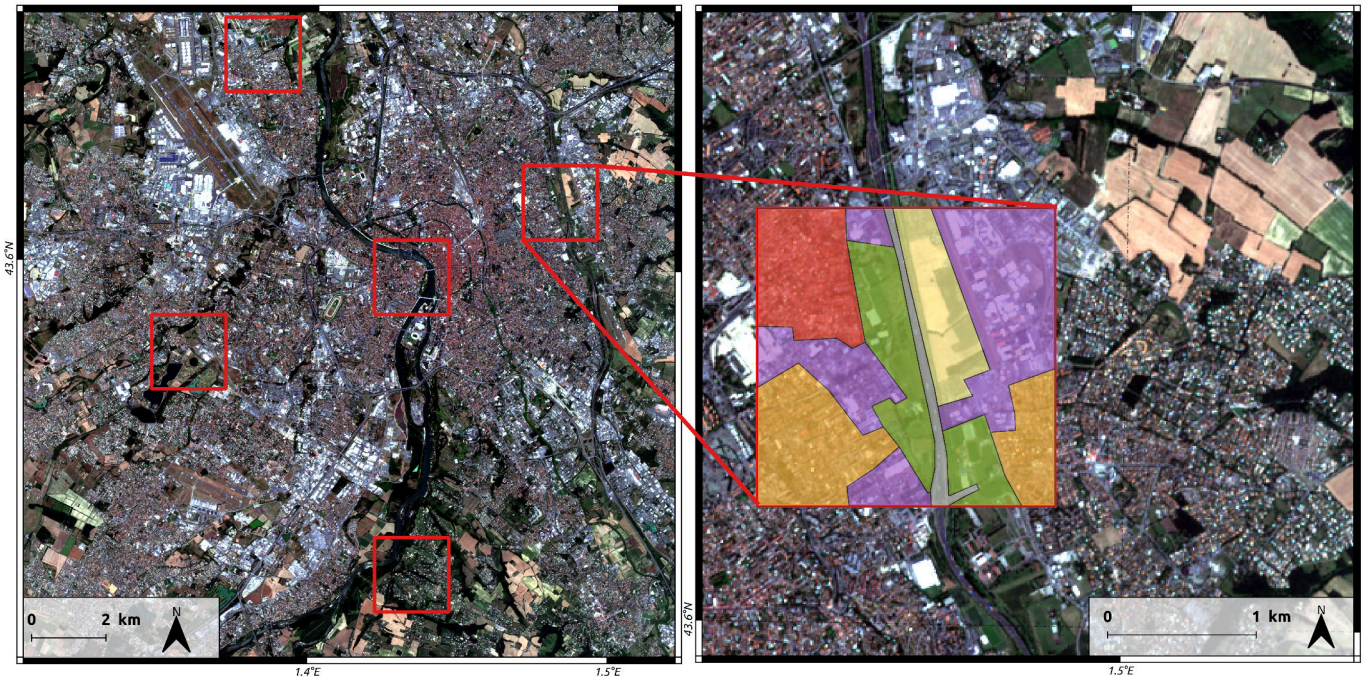
A.17 CHAPTER 6: DIGITIZATION AREAS AND RESULTS FOR RENNES
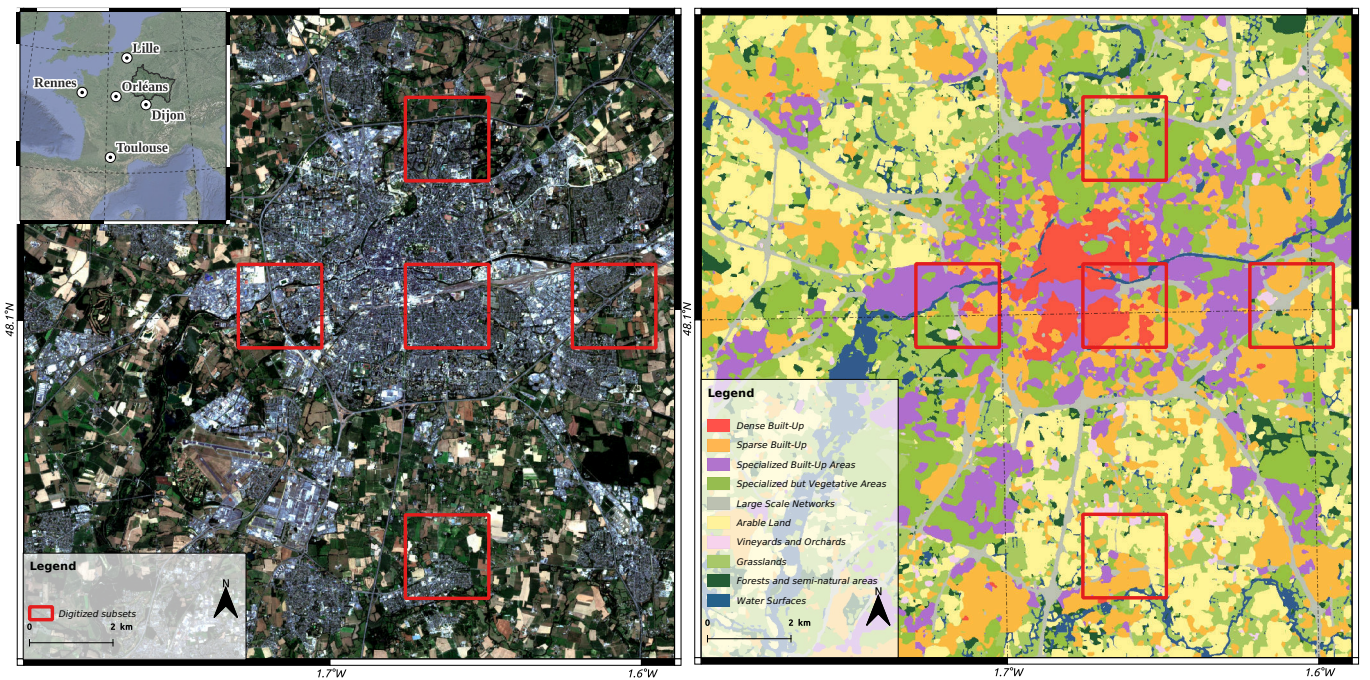
A.18   Chapter 6: Digitization areas and results for Lille
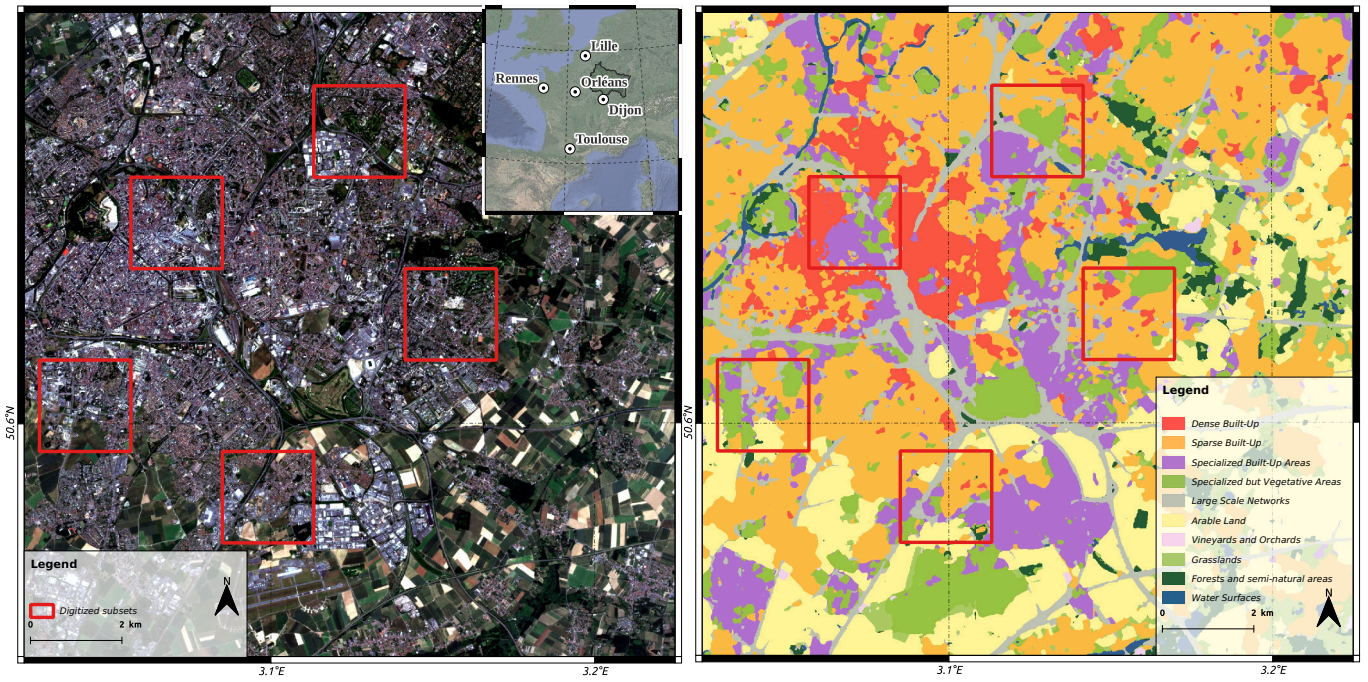
A.19   Chapter 6: Digitization areas and results for Dijon
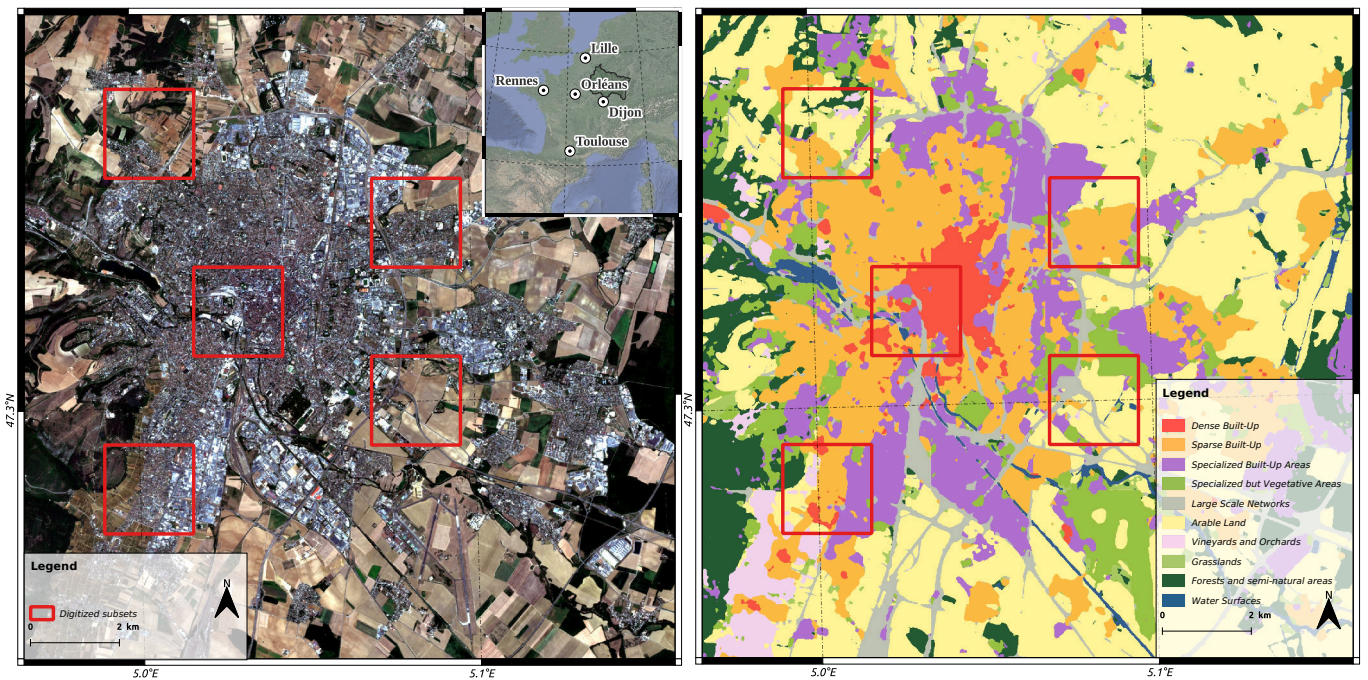
A.20 Chapter 6: Changes between natural classes and four built-up semantic classes for Châlons-en-Champagne between 2018 and 2022
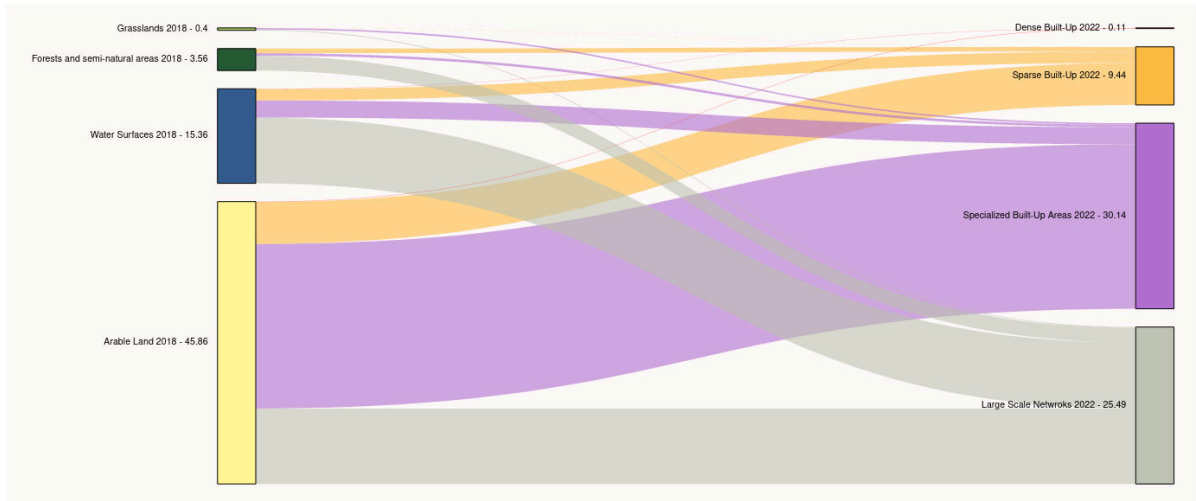
A.21 Chapter 6: Changes between natural classes and four built-up semantic classes for Metz between 2018 and 2022
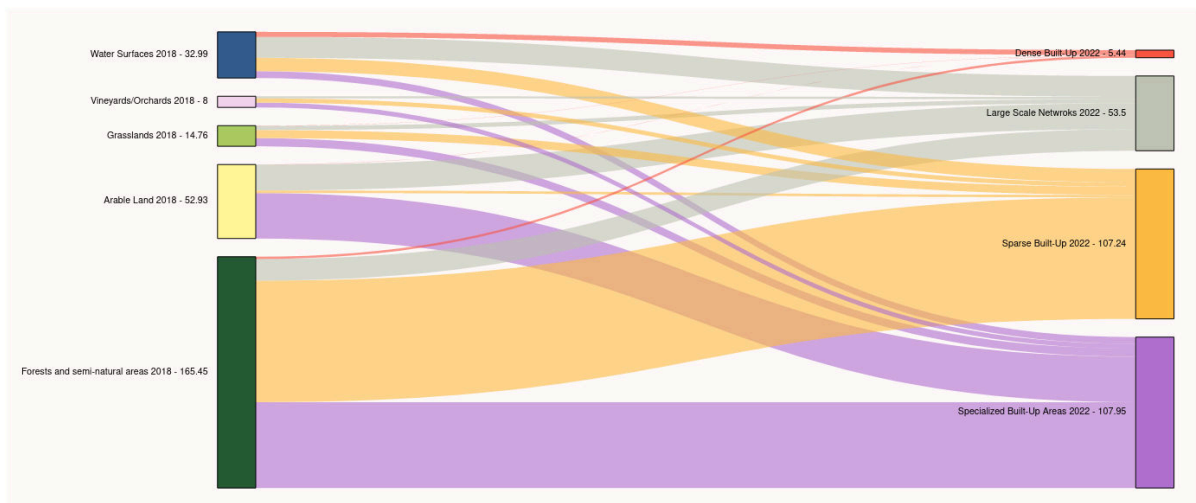
A.22 Chapter 6: Changes between natural classes and four built-up semantic classes for Mulhouse between 2018 and 2022
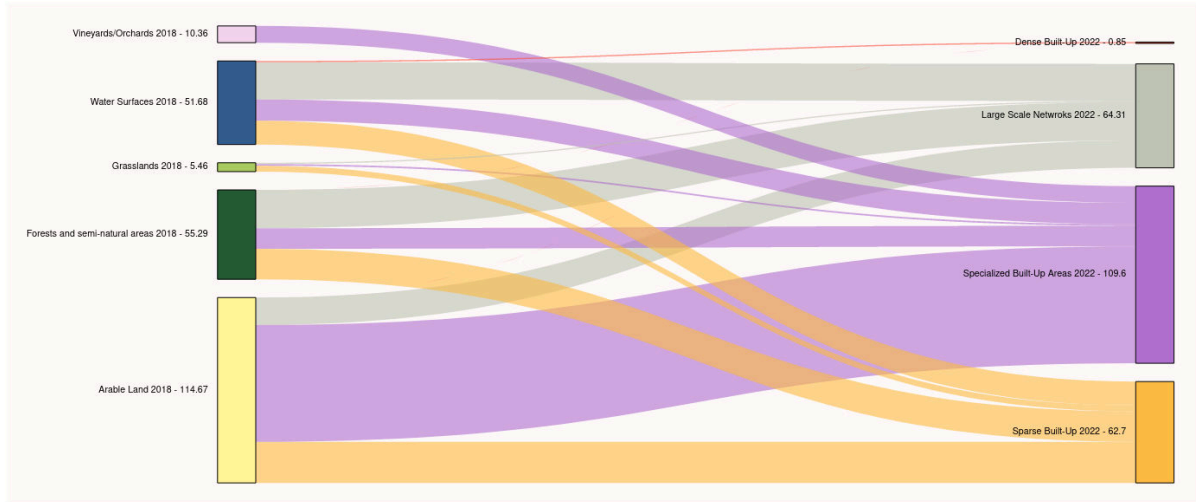
A.23 Chapter 6: Changes between natural classes and four built-up semantic classes for Nancy between 2018 and 2022
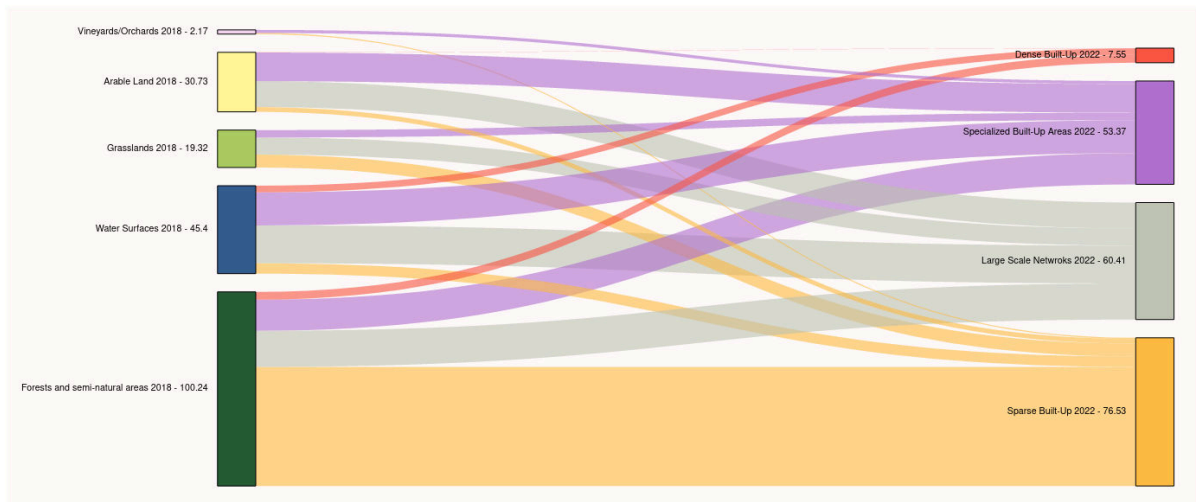
A.24 Chapter 6: Changes between natural classes and four built-up semantic classes for Reims between 2018 and 2022
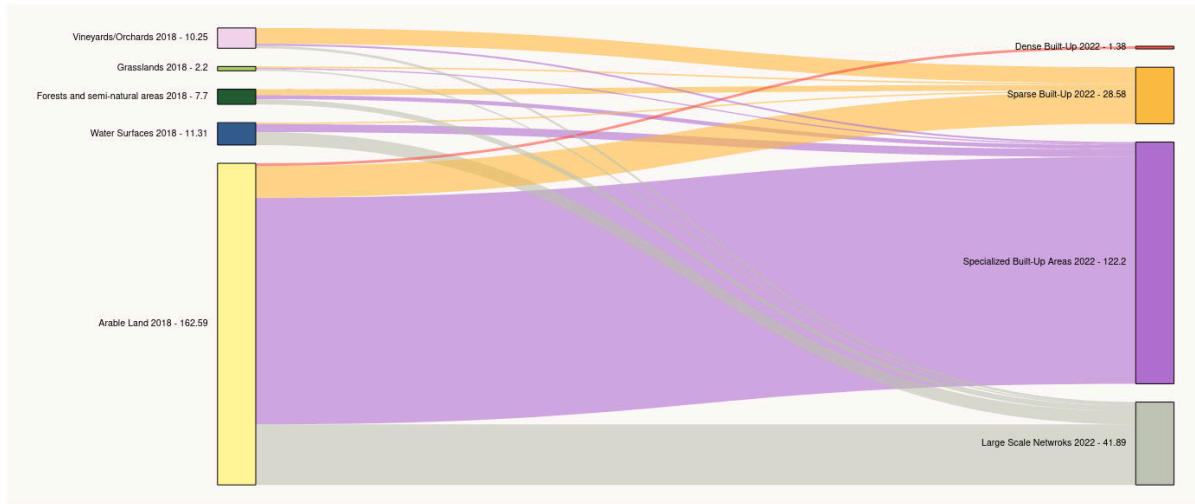
A.25 Chapter 6: Changes between natural classes and four built-up semantic classes for Strasbourg between 2018 and 2022
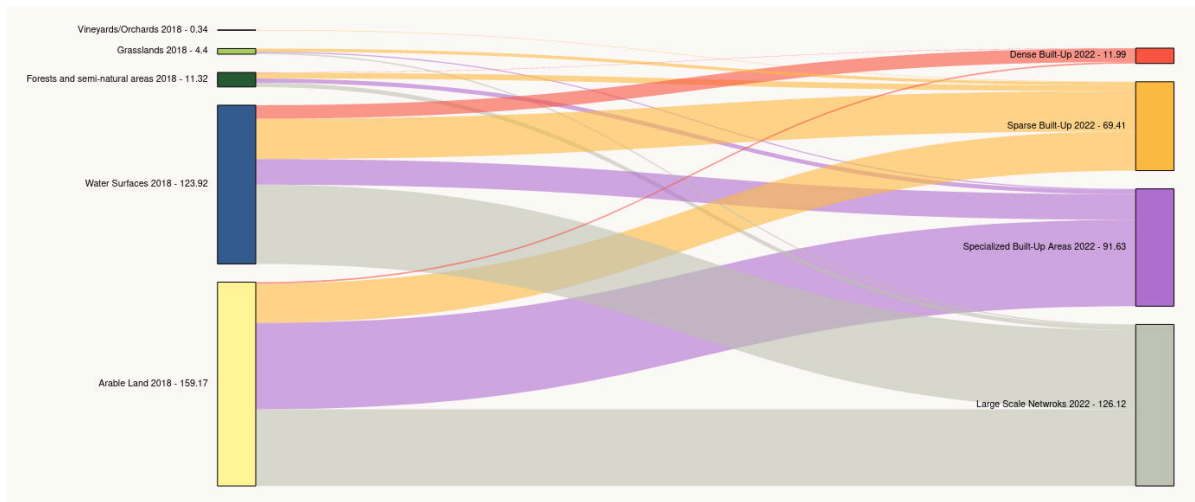
Bibliography

Abolfazl Abdollahi, Biswajeet Pradhan, Gaurav Sharma, Khairul Nizam Abdul Maulud, et Abdullah Alamri. Improving road semantic segmentation using generative adversarial network. *IEEE Access*, 9:64381–64392, 2021.

Mohammed El Amin Larabi, Souleyman Chaib, Khadidja Bakhti, et Moussa Sofiane Karoui. Transfer learning for changes detection in optical remote sensing imagery. Dans *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 1582–1585, 2019.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, et Naama Zwerdling. Do not have enough data? deep learning to the rescue! 34(05):7383–7390, 2020.

Nicolas Audebert, Bertrand Le Saux, et Sébastien Lefèvre. Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140:20–32, 2018. URL https://hal.archives-ouvertes.fr/hal-01636145.

Vijay Badrinarayanan, Alex Kendall, et Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

Haiwei Bai, Jian Cheng, Yanzhou Su, Siyu Liu, et Xin Liu. Calibrated focal loss for semantic labeling of high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:6531–6547, 2022.

Jim B Ball. Global forest resources: history and dynamics. *The Forests Handbook: An Overview of Forest Science*, 1:3–22, 2001.

Yifang Ban, Hongtao Hu, et Irene M Rangel. Fusion of quickbird ms and radarsat sar data for urban land-cover mapping: Object-based and knowledge-based approach. *International Journal of Remote Sensing*, 31(6):1391–1410, 2010.

Mariana Belgiu et Lucian Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114: 24–31, 2016.

Beatriz Bellón, Agnès Bégué, Danny Lo Seen, Claudio Aparecido De Almeida, et Margareth Simões. A remote sensing approach for regional-scale mapping of agricultural land-use systems based on ndvi time series. *Remote Sensing*, 9(6), 2017. ISSN 2072-4292.

Jon A Benediktsson, Gabriele Cavallaro, Nicola Falco, Ihsen Hedhli, Vladimir A Krylov, Gabriele Moser, Sebastiano B Serpico, et Josiane Zerubia. Remote sensing data fusion: Markov models and mathematical morphology for multisensor, multiresolution, and multiscale image classification. Dans *Mathematical Models for Remote Sensing Image Processing*, pages 277–323. Springer, 2018.

Somenath Bera et Vimal K Shrivastava. Analysis of various optimizers on deep convolutional neural network model in the application of hyperspectral remote sensing image classification. *International Journal of Remote Sensing*, 41(7):2664–2683, 2020.

Julie Betbeder, Marianne Laslier, Thomas Corpetti, Eric Pottier, Samuel Corgne, et Laurence Hubert-Moy. Multi-temporal optical and radar data fusion for crop monitoring: Application to an intensive agricultural area in brittany (france). Dans *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 1493–1496. IEEE, 2014.

S Bontemps, M Boettcher, C Brockmann, G Kirches, C Lamarche, J Radoux, M Santoro, E Vanbogaert, U Wegmüller, M Herold, et al. Multi-year global land cover mapping at 300 m and characterization for climate modelling: achievements of the land cover component of the esa climate change initiative. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(7):323, 2015.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9(1):1–17, 2022.

Agnès Bégué, Damien Arvor, Beatriz Bellon, Julie Betbeder, Diego De Abelleyra, Rodrigo P. D. Ferraz, Valentine Lebourgeois, Camille Lelong, Margareth Simões, et Santiago

R. Verón. Remote sensing and cropping practices: A review. *Remote Sensing*, 10(1), 2018. ISSN 2072-4292.

Manuel Campos-Taberner, Francisco Javier García-Haro, Beatriz Martínez, Emma Izquierdo-Verdiguier, Clement Atzberger, Gustau Camps-Valls, et María Amparo Gilabert. Understanding deep learning in land use classification based on sentinel-2 time series. *Scientific Reports*, 10(1):17188, Oct 2020. ISSN 2045-2322. URL https://doi.org/10.1038/s41598-020-74215-5.

Gavin C Cawley et Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11: 2079–2107, 2010.

Yang-Lang Chang, Tan-Hsu Tan, Tsung-Hau Chen, Joon Huang Chuah, Lena Chang, Meng-Che Wu, Narendra Babu Tatini, Shang-Chih Ma, et Mohammad Alkhaleefah. Spatial-temporal neural network for rice field classification from sar images. *Remote Sensing*, 14(8):1929, 2022.

Pats Chavez, Stuart C Sides, Jeffrey A Anderson, et al. Comparison of three different methods to merge multiresolution and multispectral data- landsat tm and spot panchromatic. *Photogrammetric Engineering and remote sensing*, 57(3):295–303, 1991.

Kaiqiang Chen, Kun Fu, Menglong Yan, Xin Gao, Xian Sun, et Xin Wei. Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(2):173–177, 2018.

Yushi Chen, Chunyang Li, Pedram Ghamisi, Xiuping Jia, et Yanfeng Gu. Deep fusion of remote sensing data for accurate classification. *IEEE Geoscience and Remote Sensing Letters*, 14(8):1253–1257, 2017.

Guillaume Chhor et Cristian Bartolome Aramburu. Satellite image segmentation for building detection using u-net. http://cs229.stanford.edu/proj2017/final-posters/5148174.pdf, 2017. [Online; accessed 20-March-2021].

François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

Nicola Clerici, Cesar Augusto Valbuena Calderón, et Juan Manuel Posada. Fusion of sentinel-1a and sentinel-2a data for land cover mapping: a case study in the lower magdalena region, colombia. *Journal of Maps*, 13(2):718–726, 2017. URL https://doi.org/10.1080/17445647.2017.1372316.

René Roland Colditz. An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sensing*, 7(8):9655–9681, 2015.

Giandomenico De Luca, João M. N. Silva, Sofia Cerasoli, João Araújo, José Campos, Salvatore Di Fazio, et Giuseppe Modica. Object-based land cover classification of cork oak woodlands using uav imagery and orfeo toolbox. *Remote Sensing*, 11(10), 2019. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/11/10/1238.

Clément Dechesne, Clément Mallet, Arnaud Le Bris, et Valérie Gouet-Brunet. Semantic segmentation of forest stands of pure species combining airborne lidar data and very high resolution multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 126:129–145, 2017.

RS DeFries et AS Belward. Global and regional land cover characterization from satellite data: an introduction to the special issue. *International Journal of Remote Sensing*, 21(6-7): 1083–1092, 2000.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, et Li Fei-Fei. Imagenet: A large-scale hierarchical image database. Dans *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Dawa Derksen, Jordi Inglada, et Julien Michel. Scaling up slic superpixels using a tile-based approach. *IEEE transactions on Geoscience and Remote Sensing*, 57(5):3073–3085, 2019.

Peng Ding, Ye Zhang, Wei-Jian Deng, Ping Jia, et Arjan Kuijper. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 141:208–218, 2018. ISSN 0924-2716. URL https://www.sciencedirect.com/science/article/pii/S0924271618301382.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, et P. Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012. ISSN 0034-4257. URL https://www.sciencedirect.com/science/article/pii/S0034425712000636. The Sentinel Missions - New Opportunities for Science.

Pauline Dusseux, Thomas Corpetti, Laurence Hubert-Moy, et Samuel Corgne. Combined use of multi-temporal optical and radar satellite images for grassland monitoring. *Remote Sensing*, 6(7):6163–6182, 2014. ISSN 2072-4292.

Lamiae El Mendili, Anne Puissant, Mehdi Chougrad, et Imane Sebari. Towards a multi-temporal deep learning approach for mapping urban fabric using sentinel 2 images. *Remote Sensing*, 12(3), 2020. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/12/3/423.

Thomas Esch, Mattia Marconcini, Andreas Felbier, Achim Roth, Wieke Heldens, Martin Huber, Max Schwinger, Hannes Taubenböck, Andreas Müller, et SJIG Dech. Urban footprint processor—fully automated processing chain generating settlement masks from global data of the tandem-x mission. *IEEE Geoscience and Remote Sensing Letters*, 10(6): 1617–1621, 2013.

Xiuqin Fang, Qiuan Zhu, Liliang Ren, Huai Chen, Kai Wang, et Changhui Peng. Large-scale detection of vegetation dynamics and their potential drivers using modis images and bfast: A case study in quebec, canada. *Remote Sensing of Environment*, 206:391–402, 2018.

Mathieu Fauvel, Jón Atli Benediktsson, Jocelyn Chanussot, et Johannes R Sveinsson. Spectral and spatial classification of hyperspectral data using svms and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3804–3814, 2008.

Federico Filipponi. Sentinel-1 grd preprocessing workflow. *Proceedings*, 18(1), 2019. ISSN 2504-3900. URL https://www.mdpi.com/2504-3900/18/1/11.

Cidália Costa Fonte, Marco Minghini, Joaquim Patriarca, Vyron Antoniou, Linda See, et Andriani Skopeliti. Generating up-to-date and detailed land use and land cover maps using openstreetmap and globeland30. *ISPRS International Journal of Geo-Information*, 6 (4):125, 2017.

Steven E Franklin, Yuhong He, Alysha Pape, Xulin Guo, et Gregory J McDermid. Landsat-comparable land cover maps using aster and spot images: a case study for large-area mapping programmes. *International journal of remote sensing*, 32(8):2185–2205, 2011.

Mark A Friedl, Douglas K McIver, John CF Hodges, Xiaoyang Y Zhang, D Muchoney, Alan H Strahler, Curtis E Woodcock, Sucharita Gopal, Annemarie Schneider, Amanda Cooper, et al. Global land cover mapping from modis: algorithms and early results. *Remote sensing of Environment*, 83(1-2):287–302, 2002.

Raffaele Gaetano, Dino Ienco, Kenji Ose, et Remi Cresson. A two-branch cnn architecture for land cover classification of pan and ms imagery. *Remote Sensing*, 10(11):1746, 2018.

Qi Gao, Mehrez Zribi, Maria Jose Escorihuela, et Nicolas Baghdadi. Synergetic use of sentinel-1 and sentinel-2 data for soil moisture mapping at 100 m resolution. *Sensors*, 17 (9), 2017. ISSN 1424-8220.

Yawogan Jean Eudes Gbodjo et Dino Ienco. Classification multi-modale de l'occupation du sol à partir de réseaux de neurones convolutifs et de l'auto-distillation. Dans *ORASIS 2021*, 2021.

France Gerard, S Petit, G Smith, A Thomson, N Brown, S Manchester, R Wadsworth, G Bugar, L Halada, P Bezak, et al. Land cover change in europe between 1950 and 2000 determined employing aerial photography. *Progress in Physical Geography*, 34(2): 183–205, 2010.

Hassan Ghassemian. A review of remote sensing image fusion methods. *Information Fusion*, 32:75–89, 2016. ISSN 1566-2535. URL https://www.sciencedirect.com/science/article/pii/S1566253516300173.

Chandra P Giri. *Remote sensing of land use and land cover: principles and applications*. CRC press, 2012.

Daniel Guidici et Matthew L Clark. One-dimensional convolutional neural network landcover classification of multi-seasonal hyperspectral imagery in the san francisco bay area, california. *Remote Sensing*, 9(6):629, 2017.

Xiaojiao Guo, Chengcai Zhang, Weiran Luo, Jing Yang, et Miao Yang. Urban impervious surface extraction based on multi-features and random forest. *IEEE Access*, 8:226609–226623, 2020.

Sebastian Hafner, Yifang Ban, et Andrea Nascetti. Unsupervised domain adaptation for global urban extraction using sentinel-1 sar and sentinel-2 msi data. *Remote Sensing of Environment*, 280:113192, 2022.

Olivier Hagolle. theia_download. https://github.com/olivierhagolle/theia_download (21 February 2020), 2021.

Chaoyi Han, Yiping Duan, Xiaoming Tao, et Jianhua Lu. Dense convolutional networks for semantic segmentation. *IEEE Access*, 7:43369–43382, 2019.

Robert M. Haralick, Its'hak Dinstein, et K. Shanmugam. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6):610–621, 1973. ISSN 21682909.

Caner Hazirbas, Lingni Ma, Csaba Domokos, et Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. Dans *Asian conference on computer vision*, pages 213–228. Springer, 2016.

Pouya Hedayati et Damian Bargiel. Fusion of sentinel-1 and sentinel-2 images for classification of agricultural areas using a novel classification approach. Dans *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 6643–6646, 2018.

Hidetake Hirayama, Ram C Sharma, Mizuki Tomita, et Keitarou Hara. Evaluating multiple classifier system for the reduction of salt-and-pepper noise in the classification of very-high-resolution satellite images. *International journal of remote sensing*, 40(7):2542–2557, 2019.

Roger LeB Hooke, José Francisco Martín Duque, et Javier de Pedraza Gilsanz. Land transformation by humans: a review. *Ene*, 12:43, 2013.

Jingliang Hu, Lichao Mou, Andreas Schmitt, et Xiao Xiang Zhu. Fusionet: A two-stream convolutional neural network for urban scene classification using polsar and hyperspectral data. Dans *2017 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4, 2017.

Bo Huang, Bei Zhao, et Yimeng Song. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 214:73–86, 2018. ISSN 0034-4257. URL https://www.sciencedirect.com/science/article/pii/S0034425718302074.

Fenghua Huang, Ying Yu, et Tinghao Feng. Automatic extraction of urban impervious surfaces based on deep learning and multi-source remote sensing data. *Journal of Visual Communication and Image Representation*, 60, 04 2019.

Wei Huang, Liang Xiao, Zhihui Wei, Hongyi Liu, et Songze Tang. A new pan-sharpening method with deep neural networks. *IEEE Geoscience and Remote Sensing Letters*, 12(5): 1037–1041, 2015.

Gianni Cristian Iannelli et Paolo Gamba. Jointly exploiting sentinel-1 and sentinel-2 for urban mapping. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 8209–8212, 2018.

Dino Ienco, Raffaele Gaetano, Claire Dupaquier, et Pierre Maurel. Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1685–1689, 2017.

Dino Ienco, Roberto Interdonato, Raffaele Gaetano, et Dinh Ho Tong Minh. Combining sentinel-1 and sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158:11–22, 2019. ISSN 0924-2716. URL https://www.sciencedirect.com/science/article/pii/S0924271619302278.

Vladimir Iglovikov et Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation, 2018.

J. Inglada, M. Arias, B. Tardy, D. Morin, S. Valero, O. Hagolle, G. Dedieu, G. Sepulcre, S. Bontemps, et P. Defourny. Benchmarking of algorithms for crop type land-cover maps using sentinel-2 image time series. Dans *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3993–3996, 2015.

Jordi Inglada, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin, et Isabel Rodes. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1), 2017. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/9/1/95.

Roberto Interdonato, Dino Ienco, Raffaele Gaetano, et Kenji Ose. Duplo: A dual view point deep learning architecture for time series classification. *ISPRS journal of photogrammetry and remote sensing*, 149:91–104, 2019.

Ahmed Iqbal, Muhammad Sharif, Muhammad Attique Khan, Wasif Nisar, et Majed Alhaisoni. Ff-unet: a u-shaped deep convolutional neural network for multimodal biomedical image segmentation. *Cognitive Computation*, 14(4):1287–1302, 2022.

Elena G. Irwin et Nancy E. Bockstael. The evolution of urban sprawl: Evidence of spatial heterogeneity and increasing land fragmentation. *Proceedings of the National Academy of Sciences*, 104(52):20672–20677, 2007. ISSN 0027-8424. URL https://www.pnas.org/content/104/52/20672.

Ilham Jamaluddin, Tipajin Thaipisutikul, Ying-Nong Chen, Chi-Hung Chuang, et Chih-Lin Hu. Mdprepost-net: A spatial-spectral-temporal fully convolutional network for mapping of mangrove degradation affected by hurricane irma 2017 using sentinel-2 data. *Remote Sensing*, 13:5042, 12 2021.

Shunping Ji, Chi Zhang, Anjian Xu, Yun Shi, et Yulin Duan. 3d convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sensing*, 10(1):75, 2018.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, et Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. Dans *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

Yongshi Jie, Xianhua Ji, Anzhi Yue, Jingbo Chen, Yupeng Deng, Jing Chen, et Yi Zhang. Combined multi-layer feature fusion and edge detection method for distributed photovoltaic power station identification. *Energies*, 13(24), 2020. ISSN 1996-1073. URL https://www.mdpi.com/1996-1073/13/24/6742.

G. A. Fotso Kamga, L Bitjoka, T. Akram, A. Mengue Mbom, S. Rameez Naqvi, et Y. Bouroubi. Advancements in satellite image classification : methodologies, techniques, approaches and applications. *International Journal of Remote Sensing*, 42(20):7662–7722, 2021. URL https://doi.org/10.1080/01431161.2021.1954261.

N. Karasiak, D. Sheeren, M. Fauvel, J. Willm, J.-F. Dejoux, et C. Monteil. Mapping tree species of forests in southwest france using sentinel-2 image time series. Dans *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pages 1–4, 2017.

Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C Mazzariello, Mark Mathis, et Steven P Brumby. Global land use/land cover with sentinel 2 and deep learning. Dans *2021 IEEE international geoscience and remote sensing symposium IGARSS*, pages 4704–4707. IEEE, 2021.

Taskin Kavzoglu et Ismail Colkesen. A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 11(5):352–359, 2009.

Ronald Kemker, Carl Salvaggio, et Christopher Kanan. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:60–77, Nov 2018. ISSN 0924-2716. URL http://dx.doi.org/10.1016/j.isprsjprs.2018.04.014.

Diederik P. Kingma et Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Thierry Koleck et Centre National des Etudes Spatiales (CNES). S1Tiling. https://gitlab.orfeo-toolbox.org/s1-tiling/s1tiling (10 September 2021), 2021.

Alex Krizhevsky, Ilya Sutskever, et Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, et Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017.

Eric F Lambin, Helmut J Geist, et Erika Lepers. Dynamics of land-use and land-cover change in tropical regions. *Annual review of environment and resources*, 28(1):205–241, 2003.

Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, et Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018.

François Leconte, Julien Bouyer, Rémy Claverie, et Mathieu Pétrissans. Estimation of spatial air temperature distribution at sub-mesoclimatic scale using the lcz scheme and mobile measurements. Dans *Proceedings of the 9th international conference on urban climate (ICUC9) jointly with 12th symposium on the urban environment, Toulouse, France*, 2015.

Yann LeCun, Yoshua Bengio, et Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Haifeng Li, Yi Li, Guo Zhang, Ruoyun Liu, Haozhe Huang, Qing Zhu, et Chao Tao. Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Jiang Li et R.M. Narayanan. A shape-based approach to change detection of lakes using time series remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 41(11):2466–2477, 2003.

Junjie Li, Yizhuo Meng, Donyu Dorjee, Xiaobing Wei, Zhiyuan Zhang, et Wen Zhang. Automatic road extraction from remote sensing imagery using ensemble learning and postprocessing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10535–10547, 2021.

Mengmeng Li, Alfred Stein, Wietske Bijker, et Qingming Zhan. Region-based urban road extraction from vhr satellite images using binary partition tree. *International Journal of Applied Earth Observation and Geoinformation*, 44:217–225, 2016.

Qingting Li, Cuizhen Wang, Bing Zhang, et Linlin Lu. Object-based crop classification with landsat-modis enhanced time-series data. *Remote Sensing*, 7(12):16091–16107, 2015. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/7/12/15820.

Ruirui Li, Wenjie Liu, Lei Yang, Shihao Sun, Wei Hu, Fan Zhang, et Wei Li. Deepunet: A deep fully convolutional network for pixel-level sea-land segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(11):3954–3962, 2018.

Ying Li, Haokui Zhang, et Qiang Shen. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing*, 9(1):67, 2017.

Jia Liu, Maoguo Gong, Kai Qin, et Puzhao Zhang. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. *IEEE Transactions on Neural Networks and Learning Systems*, 29(3):545–559, 2018.

Yansong Liu, Sankaranarayanan Piramanayagam, Sildomar T. Monteiro, et Eli Saber. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs. Dans *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1561–1570, 2017.

Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, et Guangquan Zhang. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14–23, 2015. ISSN 0950-7051. URL https://www.sciencedirect.com/science/article/pii/S0950705115000179. 25th anniversary of Knowledge-Based Systems.

Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, et Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177, 2019. ISSN 0924-2716. URL https://www.sciencedirect.com/science/article/pii/S0924271619301108.

Diego Marcos, Michele Volpi, Benjamin Kellenberger, et Devis Tuia. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:96–107, Nov 2018. ISSN 0924-2716. URL http://dx.doi.org/10.1016/j.isprsjprs.2018.01.021.

FJ Marschner. Major land uses in the united states (map scale 1: 5,000,000). *USDA Agricultural Research Service, Washington, DC*, 252, 1950.

Robert N. Masolele, Veronique De Sy, Martin Herold, Diego Marcos, Jan Verbesselt, Fabian Gieseke, Adugna G. Mullissa, et Christopher Martius. Spatial and temporal deep learning methods for deriving land-use following deforestation: A pan-tropical

case study using landsat time series. *Remote Sensing of Environment*, 264:112600, 2021. ISSN 0034-4257. URL https://www.sciencedirect.com/science/article/pii/S0034425721003205.

Aaron E. Maxwell, Timothy A. Warner, et Luis Andrés Guillén. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 1: Literature review. *Remote Sensing*, 13(13), 2021a. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/13/13/2450.

Aaron E. Maxwell, Timothy A. Warner, et Luis Andrés Guillén. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 2: Recommendations and best practices. *Remote Sensing*, 13(13), 2021b. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/13/13/2591.

Shaohui Mei, Xin Yuan, Jingyu Ji, Yifan Zhang, Shuai Wan, et Qian Du. Hyperspectral image spatial super-resolution via 3d full convolutional neural network. *Remote Sensing*, 9(11):1139, 2017.

Audrey Mercier, Julie Betbeder, Florent Rumiano, Jacques Baudry, Valéry Gond, Lilian Blanc, Clément Bourgoin, Guillaume Cornu, Carlos Ciudad, Miguel Marchamalo, et al. Evaluation of sentinel-1 and 2 time series for land cover classification of forest–agriculture mosaics in temperate and tropical landscapes. *Remote Sensing*, 11(8):979, 2019.

Pierre Merlin, Françoise Choay, et Presses Universitaires de France. Dictionnaire de l'urbanisme et de l'aménagement. 1988.

Teresa Mexia, Joana Vieira, Adriana Príncipe, Andreia Anjos, Patrícia Silva, Nuno Lopes, Catarina Freitas, Margarida Santos-Reis, Otília Correia, Cristina Branquinho, et al. Ecosystem services: Urban parks under a magnifying glass. *Environmental research*, 160:469–478, 2018.

William B Meyer et Billie L Turner. Human population growth and global land-use/cover change. *Annual review of ecology and systematics*, 23(1):39–61, 1992.

Paidamwoyo Mhangara et John Odindi. Potential of texture-based classification in urban landscapes using multispectral aerial photos. *South African Journal of Science*, 109:1–8, 12 2012.

Mthembeni Mngadi, John Odindi, Kabir Peerbhay, et Onisimo Mutanga. Examining the

effectiveness of sentinel-1 and 2 imagery for commercial forest species mapping. *Geocarto International*, 36(1):1–12, 2021.

Rahman Momeni, Paul Aplin, et Doreen S. Boyd. Mapping complex urban land cover from spaceborne imagery: The influence of spatial resolution, spectral band set and classification approach. *Remote Sensing*, 8(2), 2016. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/8/2/88.

Lichao Mou, Pedram Ghamisi, et Xiao Xiang Zhu. Fully conv-deconv network for unsupervised spectral-spatial feature extraction of hyperspectral imagery via residual learning. Dans *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5181–5184. IEEE, 2017.

Giorgos Mountrakis, Jungho Im, et Caesar Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, 2011.

Douglas M Muchoney et Barry N Haack. Change detection for monitoring forest defoliation. *Photogrammetric engineering and remote sensing*, 60(10):1243–1252, 1994.

Sina Muster, Moritz Langer, Anna Abnizova, Kathy L Young, et Julia Boike. Spatiotemporal sensitivity of modis land surface temperature anomalies indicates high potential for large-scale land cover change detection in arctic permafrost landscapes. *Remote sensing of environment*, 168:1–12, 2015.

A. K. Neves, T. S. Körting, L. M. G. Fonseca, C. D. Girolamo Neto, D. Wittich, G. A. O. P. Costa, et C. Heipke. Semantic segmentation of brazilian savanna vegetation using high spatial resolution satellite data and u-net. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3-2020:505–511, 2020. URL https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/V-3-2020/505/2020/.

Keiller Nogueira, Otávio AB Penatti, et Jefersson A Dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539–556, 2017.

VAO Odenyo et DE Pettry. Land-use mapping by machine processing of landsat-1 data. *Photogrammetric Engineering and Remote Sensing*, 43(4), 1977.

Maxime Oquab, Leon Bottou, Ivan Laptev, et Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. Dans *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, 2014.

Song Ouyang et Yansheng Li. Combining deep semantic segmentation network and graph convolutional neural network for semantic segmentation of remote sensing imagery. *Remote Sensing*, 13(1), 2021. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/13/1/119.

Ana María Pacheco-Pascagaza, Yaqing Gou, Valentin Louis, John F Roberts, Pedro Rodríguez-Veiga, Polyanna da Conceição Bispo, Fernando DB Espírito-Santo, Ciaran Robb, Caroline Upton, Gustavo Galindo, et al. Near real-time change detection system using sentinel-2 and machine learning: a test for mexican and colombian forests. *Remote Sensing*, 14(3):707, 2022.

Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005.

Antigoni Panagiotopoulou, Lazaros Grammatikopoulos, Georgia Kalousi, et Eleni Charou. Sentinel-2 and spot-7 images in machine learning frameworks for super-resolution. Dans *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*, pages 462–476. Springer, 2021.

Maria Papadomanolaki, Maria Vakalopoulou, et Konstantinos Karantzalos. A novel object-based deep learning framework for semantic segmentation of very high-resolution remote sensing data: Comparison with convolutional and fully convolutional networks. *Remote Sensing*, 11(6):684, 2019a.

Maria Papadomanolaki, Sagar Verma, Maria Vakalopoulou, Siddharth Gupta, et Konstantinos Karantzalos. Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data. Dans *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 214–217. IEEE, 2019b.

Charlotte Pelletier, Geoffrey I. Webb, et François Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5), 2019. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/11/5/523.

Andreas Pfeuffer, Karina Schulz, et Klaus Dietmayer. Semantic segmentation of video sequences with convolutional lstms. Dans *2019 IEEE intelligent vehicles symposium (IV)*, pages 1441–1447. IEEE, 2019.

Paul D. Pickell, Txomin Hermosilla, Ryan J. Frazier, Nicholas C. Coops, et Michael A. Wulder. Forest recovery trends derived from landsat time series for north american boreal forests. *International Journal of Remote Sensing*, 37(1):138–149, 2016.

Sankaranarayanan Piramanayagam, Eli Saber, Wade Schwartzkopf, et Frederick W Koehler. Supervised classification of multisensor remotely sensed images using a deep learning framework. *Remote sensing*, 10(9):1429, 2018.

Rafael Pires de Lima et Kurt Marfurt. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sensing*, 12(1):86, 2019.

Cle Pohl et John L Van Genderen. Review article multisensor image fusion in remote sensing: concepts, methods and applications. *International journal of remote sensing*, 19(5): 823–854, 1998.

Tristan Postadjian, Arnaud Le Bris, Hichem Sahbi, et Clément Mallet. Investigating the potential of deep neural networks for large-scale classification of very high resolution satellite images. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2017.

Salvatore Praticò, Francesco Solano, Salvatore Di Fazio, et Giuseppe Modica. Machine learning classification of mediterranean forest habitats in google earth engine based on seasonal sentinel-2 time-series and input image composition optimisation. *Remote Sensing*, 13(4), 2021. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/13/4/586.

Anne Puissant, Nicolas Lachiche, Grzegorz Skupinski, Agnès Braud, Julien Perret, et Annabelle Mas. Classification et évolution des tissus urbains à partir de données vectorielles. *Revue internationale de géomatique*, 21:513–532, 12 2011.

Anne Puissant, Simon Rougier, et André Stumpf. Object-oriented mapping of urban trees using random forest classifiers. *International Journal of Applied Earth Observation and Geoinformation*, 26:235–245, 2014.

Abdullah F Rahman, Danilo Dragoni, Kamel Didan, Armando Barreto-Munoz, et Joseph A Hutabarat. Detecting large scale conversion of mangroves to aquaculture with change point and mixed-pixel analyses of high-fidelity modis data. *Remote Sensing of Environment*, 130:96–107, 2013.

Sivaramakrishnan Rajaraman, Sudhir Sornapudi, Philip O Alderson, Les R Folio, et Sameer K Antani. Interpreting deep ensemble learning through radiologist anno-

tations for covid-19 detection in chest radiographs. *medRxiv*, 2020. URL https://www.medrxiv.org/content/early/2020/07/16/2020.07.15.20154385.

Alexander Rakhlin, Alex Davydow, et Sergey Nikolenko. Land cover classification from satellite imagery with u-net and lovász-softmax loss. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 262–266, 2018.

Robert H Rogers, JB McKeon, LE Reed, NF Schmidt, et Roger N Schecter. Computer mapping of landsat data for environmental applications. Dans *Workshop for Environ. Applications of Multispectral Imagery*, numéro NASA-CR-145756, 1975.

Olaf Ronneberger, Philipp Fischer, et Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Simon Rougier, Anne Puissant, André Stumpf, et Nicolas Lachiche. Comparison of sampling strategies for object-based classification of urban vegetation from very high resolution satellite images. *International Journal of Applied Earth Observation and Geoinformation*, 51:60–73, 2016.

J.W. Rouse. Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation. *NASA/GSFC Type III Final report*, 1973.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, et Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. ISSN 15731405. URL http://dx.doi.org/10.1007/s11263-015-0816-y.

Marc Rußwurm et Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4):129, 2018.

Marc Rußwurm, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre, et Marco Körner. Breizhcrops: A time series dataset for crop type mapping. *arXiv preprint arXiv:1905.11893*, 2019.

Marc Rußwurm, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre, et Marco Körner. Breizhcrops: A time series dataset for crop type mapping. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences ISPRS (2020)*, 2020.

Marciano Saraiva, Églen Protas, Moisés Salgado, et Carlos Souza. Automatic mapping of center pivot irrigation systems from satellite images using deep learning. *Remote Sensing*, 12(3), 2020. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/12/3/558.

M Schiavina, M Melchiorri, M Pesaresi, P Politis, S Freire, L Maffenini, P Florio, D Ehrlich, K Goch, P Tommasi, et al. Ghsl data package 2022: public release ghs p2022, 2022.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, et Xiao Xiang Zhu. Sen12ms – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion, 2019a.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, et Xiao Xiang Zhu. SEN12MS - A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *CoRR*, abs/1906.07789, 2019b. URL http://arxiv.org/abs/1906.07789.

Michael Schmitt et Xiao Xiang Zhu. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4):6–23, 2016.

Oliver Sefrin, Felix M. Riese, et Sina Keller. Deep learning for land cover change detection. *Remote Sensing*, 13(1), 2021. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/13/1/78.

Yuri Shendryk, Yannik Rist, Catherine Ticehurst, et Peter Thorburn. Deep learning for multi-modal classification of cloud, shadow and land cover scenes in planetscope and sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157:124–136, 2019. ISSN 0924-2716. URL https://www.sciencedirect.com/science/article/pii/S0924271619302023.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, et Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *CoRR*, abs/1506.04214, 2015. URL http://arxiv.org/abs/1506.04214.

Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, et Ronald M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

Amanpreet Singh, Narina Thakur, et Aakanksha Sharma. A review of supervised machine

learning algorithms. Dans *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee, 2016.

Philippe Smets et al. What is dempster-shafer's model. *Advances in the Dempster-Shafer theory of evidence*, 34, 1994.

Yang Song, Zhifei Zhang, Razieh Kaviani Baghbaderani, Fanqi Wang, Ying Qu, Craig Stuttsy, et Hairong Qi. Land cover classification for satellite images through 1d cnn. Dans *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5. IEEE, 2019.

Pedro Walfir M. Souza-Filho, Wilson R. Nascimento, Diogo C. Santos, Eliseu J. Weber, Renato O. Silva, et José O. Siqueira. A geobia approach for multitemporal land-cover and land-use change analysis in a tropical watershed in the southeastern amazon. *Remote Sensing*, 10(11), 2018. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/10/11/1683.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, et Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Jan Stefanski, Benjamin Mack, et Björn Waske. Optimization of object-based image analysis with random forests for land cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(6):2492–2504, 2013.

Max J. Steinhausen, Paul D. Wagner, Balaji Narasimhan, et Björn Waske. Combining sentinel-1 and sentinel-2 data for improved land use and land cover mapping of monsoon regions. *International Journal of Applied Earth Observation and Geoinformation*, 73: 595–604, 2018. ISSN 0303-2434.

Wei Su, Jing Li, Yun Chen, Zhigang Liu, Jinshui Zhang, Tsuey Low, Inbaraj Suppiah, et Atikah Hashim. Textural and local spatial statistics for the object-oriented classification of urban areas using high resolution imagery. *International Journal of Remote Sensing - INT J REMOTE SENS*, 29:3105–3117, 06 2008.

Gencer Sumbul, Marcela Charfuelan, Begum Demir, et Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Jul 2019. URL http://dx.doi.org/10.1109/IGARSS.2019.8900532.

Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begum Demir, et Volker Markl. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9 (3):174–180, Sep 2021. ISSN 2373-7468. URL http://dx.doi.org/10.1109/MGRS.2021.3089174.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, et Andrew Rabinovich. Going deeper with convolutions. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

William J Todd. Urban and regional land use change detected by using landsat data. *Journal of Research of the US Geological Survey*, 5(5):529–534, 1977.

Olga Troyanskaya, Mike Cantor, Gavin Sherlock, Trevor Hastie, Rob Tibshirani, David Botstein, et Russ Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17:520–525, 07 2001.

Kabir Uddin, Mir Abdul Matin, et Sajana Maharjan. Assessment of land cover change and its impact on changes in soil erosion risk in nepal. *Sustainability*, 10(12), 2018. ISSN 2071-1050. URL https://www.mdpi.com/2071-1050/10/12/4715.

United Nations Department of Economic and Social Affairs Population Division. The World's Cities in 2018. https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf, 2018a. [Online; accessed 09-February-2021].

United Nations Department of Economic and Social Affairs Population Division. The World's Cities in 2018. https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf, 2018b. [Online; accessed 09-February-2021].

Dongjie Wang, Yan Yang, et Shangming Ning. Deepstcl: A deep spatio-temporal convlstm for travel demand prediction. Dans *2018 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2018.

Pengliang Wei, Dengfeng Chai, Tao Lin, Chao Tang, Meiqi Du, et Jingfeng Huang. Large-scale rice mapping under different years based on time-series sentinel-1 images using deep semantic segmentation model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:198–214, 2021. ISSN 0924-2716. URL https://www.sciencedirect.com/science/article/pii/S0924271621000502.

R Welch, Clifton W Pannell, et CP Lo. Land use in northeast china, 1973: A view from landsat-1. *Annals of the Association of American Geographers*, 65(4):595–596, 1975.

R. Wenger, A. Puissant, et D. Michéa. Towards an annual urban settlement map in france at 10m spatial resolution using a method for massive streams of sentinel-2. *LIVE CNRS UMR7362, Strasbourg, France*, (to be submitted), 2023a.

R. Wenger, A. Puissant, J. Weber, L. Idoumghar, et G. Forestier. Multisenge: A multimodal and multitemporal benchmark dataset for land use/land cover remote sensing applications. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3-2022:635–640, 2022a. URL https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/V-3-2022/635/2022/.

Romain Wenger, Anne Puissant, Jonathan Weber, Lhassane Idoumghar, et Germain Forestier. A new remote sensing benchmark dataset for machine learning applications : MultiSenGE, Mars 2022b. URL https://doi.org/10.5281/zenodo.6375466. ANR-17-CE23-0015.

Romain Wenger, Anne Puissant, Jonathan Weber, Lhassane Idoumghar, et Germain Forestier. U-net feature fusion for multi-class semantic segmentation of urban fabrics from sentinel-2 imagery: an application on grand est region, france. *International Journal of Remote Sensing*, 43(6):1983–2011, 2022c. URL https://doi.org/10.1080/01431161.2022.2054295.

Romain Wenger, Anne Puissant, Jonathan Weber, Lhassane Idoumghar, et Germain Forestier. Multimodal and multitemporal land use/land cover semantic segmentation on sentinel-1 and sentinel-2 imagery: An application on a multisenge dataset. *Remote Sensing*, 15(1), 2023b. ISSN 2072-4292. URL https://www.mdpi.com/2072-4292/15/1/151.

Michael A. Wulder, Jeffrey G. Masek, Warren B. Cohen, Thomas R. Loveland, et Curtis E. Woodcock. Opening the archive: How free data has enabled the science and monitoring promise of landsat. *Remote Sensing of Environment*, 122:2–10, 2012. ISSN 0034-4257. Landsat Legacy Special Issue.

Michael Wurm, Thomas Stark, Xiao Xiang Zhu, Matthias Weigand, et Hannes Taubenböck. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:

59–69, 2019. ISSN 0924-2716. URL https://www.sciencedirect.com/science/article/pii/S0924271619300383.

Michael Xie, Neal Jean, Marshall Burke, David Lobell, et Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping, 2016.

Yinghui Xing, Min Wang, Shuyuan Yang, et Licheng Jiao. Pan-sharpening via deep metric learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:165–183, 2018. ISSN 0924-2716. URL https://www.sciencedirect.com/science/article/pii/S0924271618300212. Deep Learning RS Data.

Xiaodong Xu, Wei Li, Qiong Ran, Qian Du, Lianru Gao, et Bing Zhang. Multisource remote sensing data classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):937–949, 2017.

Pavel Yakubovskiy. Segmentation models. https://github.com/qubvel/segmentation_models, 2019.

Xinyan Yang, Yuguo Li, Zhiwen Luo, et Pak Wai Chan. The urban cool island phenomenon in a high-rise high-density city and its mechanisms. *International Journal of Climatology*, 37(2):890–904, 2017. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.4747.

Zhao Yi et Xu Jianhui. Impervious surface extraction with linear spectral mixture analysis integrating principal components analysis and normalized difference building index. Dans *2016 4th International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, pages 428–432, 2016.

Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, et Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018.

Daniele Zanaga, Ruben Van De Kerchove, Wanda De Keersmaecker, Niels Souverijns, Carsten Brockmann, Ralf Quast, Jan Wevers, Alex Grosu, Audrey Paccini, Sylvain Vergnaud, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Ivelina Georgieva, Myroslava Lesiv, Sarah Carter, Martin Herold, Linlin Li, Nandin-Erdene Tsendbazar, Fabrizio Ramoino, et Olivier Arino. Esa worldcover 10 m 2020 v100, Octobre 2021. URL https://doi.org/10.5281/zenodo.5571936.

Yong Zha, Jingqing Gao, et S. Ni. Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International Journal of Remote Sensing - INT J REMOTE SENS*, 24:583–594, 02 2003.

Chuanrong Zhang et Xinba Li. Land use and land cover mapping in the era of big data. *Land*, 11(10):1692, 2022.

Haokui Zhang, Ying Li, Yuzhu Zhang, et Qiang Shen. Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. *Remote sensing letters*, 8(5):438–447, 2017.

Jixian Zhang. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, 1(1):5–24, 2010.

Pengbin Zhang, Yinghai Ke, Zhenxin Zhang, Mingli Wang, Peng Li, et Shuangyue Zhang. Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery. *Sensors*, 18(11):3717, 2018.

Puzhao Zhang, Yifang Ban, et Andrea Nascetti. Learning u-net without forgetting for near real-time wildfire monitoring by the fusion of sar and optical time series. *Remote Sensing of Environment*, 261:112467, 2021. ISSN 0034-4257. URL https://www.sciencedirect.com/science/article/pii/S0034425721001851.

Yanfei Zhong, Xiaobing Han, et Liangpei Zhang. Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 138:281–294, 2018. ISSN 0924-2716. URL https://www.sciencedirect.com/science/article/pii/S0924271618300492.

Tao Zhou, Meifang Zhao, Chuanliang Sun, et Jianjun Pan. Exploring the impact of seasonality on urban land-cover mapping using multi-season sentinel-1a and gf-1 wfv images in a subtropical monsoon-climate region. *ISPRS International Journal of Geo-Information*, 7(1), 2018. ISSN 2220-9964. URL https://www.mdpi.com/2220-9964/7/1/3.

Lin Zhu, Yushi Chen, Pedram Ghamisi, et Jón Atli Benediktsson. Generative adversarial networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9):5046–5063, 2018.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, et Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017a.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, et Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017b.

Zhe Zhu, Yuyu Zhou, Karen C. Seto, Eleanor C. Stokes, Chengbin Deng, Steward T.A. Pickett, et Hannes Taubenböck. Understanding an urbanizing planet: Strategic directions for remote sensing. *Remote Sensing of Environment*, 228:164–182, 2019. ISSN 0034-4257. URL https://www.sciencedirect.com/science/article/pii/S0034425719301658.

Romain Wenger

Apport des images Sentinel-1&2 et des méthodes d'apprentissage profond pour la cartographie et le suivi des modes d'occupation des sols

Résumé

La couverture biophysique de la surface terrestre, plus communément appelée occupation des sols, est une composant des Variables Climatiques Essentielles (VCE) depuis plusieurs années. Dans un contexte de changement globale, sa connaissance précise est au cœur de nombreux projets de recherche sur l'environnement. En effet, cette donnée peut être utilisé en entrée de différents modèles, tels que les modèles climatiques et permet d'aider les acteurs locaux des territoires dans leur prise de décision. Les constellations satellitaires actuelles fournissent des images à haute résolution spatiale et temporelle avec de multiples modes d'acquisition (radar, optique etc.). Les nouvelles méthodes d'intelligence artificielle telles que l'apprentissage profond, ont permis d'obtenir des résultats prometteurs et ont accéléré le traitement des données massives. Ainsi, l'objectif de cette thèse est d'évaluer la contribution des capteurs Sentinel-1&2 pour la classification et le suivi de l'occupation et de l'usage des sols et plus particulièrement des tissus urbains.

Mots clefs : télédétection – intelligence artificielle – apprentissage profond – Sentinel-1 – Sentinel-2

Résumé en anglais

The biophysical cover of the earth's surface, otherwise known as land cover, has been an integral part of Essential Climate Variables (ECV) for several years. In a context of global change, precise knowledge of land cover is an important component of many environmental research projects. Indeed, this data can be used as input for several types of models (e.g. climate models) and help urban planners in their decision making. For several years, constellations of Earth observation satellites have been able to transmit high spatial and temporal resolution images (e.g. Sentinel constellation) using several acquisition modes (radar, optical, etc.). Deep learning methods, which have become increasingly attractive in recent years, have accelerated the processing of these large volumes of data and have also produced better results than traditional machine learning methods. The objective of this thesis is to evaluate the contribution of Sentinel-1&2 sensors for the classification and monitoring of land use/land cover and more specifically of urban fabrics.

Keywords : remote sensing – artificial intelligence – deep learning – Sentinel-1 – Sentinel-2