Assessing Data Volume Redundancy in the Sentinel-2 Tiling Grid System

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

066 421 Geodäsie und Geoinformation

eingereicht von

Konstantin Falkner, BSc
Matrikelnummer 11910550

ausgeführt am Department für Geodäsie und Geoinformation
der Fakultät für Mathematik und Geoinformation
der Technischen Universität Wien

Betreuung:
Univ.Prof. Dipl.-Ing. Dr.techn. Wolfgang Wagner
Senior Scientist Dipl.-Ing. Dr.techn. Bernhard Bauer-Marschallinger

Wien, 23. Juni 2025

Konstantin Falkner	Wolfgang Wagner

Assessing Data Volume Redundancy in the Sentinel-2 Tiling Grid System

DIPLOMA THESIS

submitted in partial fulfilment of the requirements for the degree of

Diplom-Ingenieur

in

066 421 Geodäsie und Geoinformation

by

Konstantin Falkner, Bsc
Registration Number 11910550

at the Department of Geodesy und Geoinformation
of the Faculty of Mathematics and Geoinformation
at TU Wien

Supervision:
Univ.Prof. Dipl.-Ing. Dr.techn. Wolfgang Wagner
Senior Scientist Dipl.-Ing. Dr.techn. Bernhard Bauer-Marschallinger

Vienna, June 23rd, 2025

Konstantin Falkner

Wolfgang Wagner

Declaration

I hereby declare that I have written this thesis independently and have not used any sources or aids other than those specified. Writing assistance software has been utilised solely as a tool to enhance and verify my own writing style. This thesis has not been presented in any form to any other testing authority before, either in Austria or abroad.

Vienna, June 23rd, 2025

Konstantin Falkner

Abstract

Grid-based spatial reference systems serve as fundamental frameworks for the efficient handling of Earth Observation (EO) raster imagery. By projecting and partitioning the Earth's surface onto regular planar grids and tiles, such Tiling Grid Systems (TGS) enable consistent georeferencing, simplify the management of large datasets, and support the creation of spatio-temporal datacubes. This facilitates efficient access, processing, and integration of EO data across different timeframes and sensors. The Sentinel-2 (S-2) mission employs a TGS based on the Universal Transverse Mercator projections and Military Grid Reference System tiles. However, its design results in significant overlapping tile coverage, causing regions of the Earth's surface to be mapped multiple times. This redundancy leads to data duplication, increasing storage, processing, and dissemination demands, which ultimately reduces the efficiency of S-2 imagery usage. The objective of this thesis is to quantify the extent of this redundancy across global landmasses using S-2 datasets. By assessing the volumes of the datasets stored in both the S-2 TGS, including data duplicates, and in a reference framework nearly free of redundancy, the system's efficiency is evaluated through direct comparison. Datasets with varying properties and different experimental setups are employed to investigate how redundancy depends on the type of data as well as geographic regions. The results align with previous studies based on land surface area calculations and reveal that the S-2 TGS causes substantial overheads of up to 32% additional data volume. These findings underscore the necessity of optimised spatial reference frameworks to support the increasing requirements of modern EO data.

Kurzfassung

Gitterbasierte räumliche Referenzsysteme bilden die Grundlage für den effizienten Umgang mit Rasterdaten der Erdbeobachtung. Durch die Abbildung und Unterteilung der Erdoberfläche in regelmäßige, planare Gitter und Kacheln ermöglichen sogenannte Tiling Grid Systems (TGS) eine einheitliche Georeferenzierung, vereinfachen die Verwaltung umfangreicher Datensätze und unterstützen die Erstellung raum-zeitlicher Datenwürfel. Dadurch wird der effiziente Zugriff auf Erdbeobachtungsdaten sowie deren Verarbeitung und Integration über verschiedene Zeiträume und Sensorsysteme hinweg erleichtert. Die Sentinel-2 (S-2) Mission nutzt ein solches TGS, das auf den Universal Transverse Mercator Projektionen und Kacheln des Military Grid Reference Systems basiert. Allerdings führt dieses Design zu einer erheblichen Überlappung von Kacheln, wodurch bestimmte Regionen der Erdoberfläche mehrfach abgebildet werden. Diese Redundanz erzeugt Datenduplikate, was den Bedarf an Speicherplatz, Rechenressourcen und Übertragungskapazitäten erhöht und letztlich die Gesamteffizienz der Nutzung von S-2 Daten verringert. Diese Arbeit quantifiziert die durch das S-2 TGS verursachten Redundanzen über globalen Landmassen anhand von S-2 Datensätzen. Dazu werden die Datensätze sowohl im S-2 TGS also inklusive Redundanzen als auch in einem nahezu redundanzfreien Referenzsystem Redundanzen gespeichert. Durch den Vergleich der dabei entstehenden Datenvolumina wird die Effizienz des Systems bewertet. Verschiedene Datensätze mit unterschiedlichen Eigenschaften sowie mehrere Analyseszenarien werden eingesetzt, um den Einfluss von Datentypen und geografischen Regionen auf das Ausmaß der Redundanz zu untersuchen. Die Ergebnisse bestätigen frühere Studien basierend auf Landflächenberechnungen und zeigen, dass das S-2 TGS zu einem erheblichen Mehraufwand von bis zu 32% an zusätzlichem Datenvolumen führen kann. Diese Erkenntnisse verdeutlichen die Notwendigkeit optimierter räumlicher Referenzsysteme, um den steigenden Anforderungen an die Speicherung, Verarbeitung und Bereitstellung moderner Erdbeobachtungsdaten gerecht zu werden.

Contents

List of Figures

List of Tables

Abbreviations

ARD Analysis Ready Data

CONUS Contiguous United States

CRS Coordinate Reference System

E7G Equi7Grid

EPSG European Petroleum Survey Group

GDAL Geospatial Data Abstraction Library

L2A Level-2A data

MGRS Military Grid Reference System

RGB Red-Green-Blue

S-2 Sentinel-2

S2M S2MOSAIC L2A 2017 collection

SCL Scene Classification Layer

SR Surface Reflectance

TG Tiling Grid

TGS Tiling Grid System

TIFF Tagged Image File Format

UPS Universal Polar Stereographic (System)

UTM Universal Transverse Mercator (System)

WGS84 World Geodetic System 1984

WRS-2 World Reference System 2

1 Introduction

Remote sensing satellites equipped with imaging instruments systematically scan the Earth's surface and capture data in the form of images. These images are essentially two-dimensional numeric arrays in native sensor geometry, containing raw measurements. Each measurement represents the aggregated energy over the footprint of each pixel on the ground (Lillesand et al., 2015). To make an image usable, it must undergo various geometric and radiometric corrections and has to be registered to a common spatial framework. This georeferencing process allows for the combination of multiple images or integration with data from other sources (Richards, 2022).

As a pivotal system for processing, storing and further disseminating data, such a framework should directly comply with the discrete raster format of imagery while representing the Earth's surface with utmost accuracy and efficiency. The foundation of the aforementioned system is the definition of an ellipsoid, which is positioned and oriented by a geodetic datum to approximate the real Earth. The resulting geographic coordinate reference system (CRS) is then transformed into a projected CRS by applying a map projection, which converts ellipsoidal coordinates to planar, Cartesian coordinates (ISO 19111:2019, 2019). Discretising this continuous two-dimensional space with the help of a grid definition allows for the consistent reference of information to coordinates. Although grids can be irregular and of arbitrary dimensions, regular two-dimensional grids with equally spaced and axis-parallel grid lines are typically used for imagery, as these grids form arrays of square cells (ISO 19123-1:2023, 2023). Mapping image pixels from sensor geometry onto these grid cells accomplishes registration to the underlying CRS while preserving the native raster format, as pixels correspond to the cells of a regular grid. This type of grid is referred to as *pixel grid* (*PG*) in this thesis.

When defining a PG with a broad spatial extent at an arbitrarily high resolution, processing and storing it may be limited by computer memory or file format restrictions. Furthermore, accessing only specific regions is often necessary in practical applications. Therefore, the PG extent is partitioned into tiles, containing only manageable amounts of adjacent cells. For the partition of regular grids, tile matrices are used, which act like an additional, higher-order regular grid that organises the cells of the PG into tiles (ISO/DIS 19177-1, 2024). This concept is referred to as *tiling grid* (*TG*) in this thesis. It produces a set of smaller arrays that can be stored and accessed as independent files, facilitating the processing of large gridded datasets. Additionally, it enables pixel coordinates to be efficiently determined through array internal indexing, using just a minimal set of parameters to define each tile (Kopp et al., 2019).

A TG fixed to the Earth's surface also meets the Analysis Ready Data (ARD) concept outlined by the Committee on Earth Observation Satellites. By serving as a unified spatial frame-

work, it ensures compatibility across datasets and enables the creation of data cubes from geospatial imagery. Regardless of acquisition geometries or types of sensors, entire scenes can be mapped onto the predefined TG. This allows for stacking into a data cube, facilitating multi-temporal and multi-sensor analyses (Lewis et al., 2018).

However, regular TGs feature a significant drawback due to cartographic mapping functions. These functions inevitably introduce geometric distortions during the projection from ellipsoidal to planar coordinates. Depending on the type of map projection, either lengths, areas, or angles are preserved without distortion, but never all three. Choosing an appropriate projection depends heavily on the application and intended use of the data. For instance, projections that preserve area are widely used for thematic mapping where accurate representation of spatial distributions is essential. In contrast, conformal projections are preferred for navigation purposes due to their ability to preserve local angles and shapes. Conversely, these projections pose challenges for applications that require accurate distance measurements (Snyder, 1987). For discrete, gridded data, distortion results in pixel loss and pixel duplication as under- and oversampling occurs during the mapping onto a PG. Consequently, selecting optimal projection methods and parameters is crucial for maintaining accuracy and efficiency (Mulcahy, 2000; White, 2006).

Limiting a TG's coverage to a regional extent minimises distortion, as the projection parameters can be optimised around a particular central point or meridian (Finn et al., 2012). By combining several TGs, global coverage can be achieved nonetheless. However, each TG operates independently, as each projection maps onto its own uniquely defined CRS. This approach, referred to in this work as *tiling grid system* (*TGS*), allows the handling of data at a global scale without introducing significant geometric distortion. However, using multiple TGs instead of a single one introduces increased complexity.

On one hand, this approach challenges the fundamental principle of a unified reference framework for all data. For instance, data that spans more than one TG is not referenced within the same spatial domain and, as a result, cannot be directly used. Addressing this requires reprojecting and resampling the data to a common system, which compromises its geometric accuracy (Roy et al., 2016; Yan et al., 2018).

On the other hand, combining multiple TGs inevitably results in data duplication. The area that must be covered by the tiles of a TG is determined by the geographic extent of its underlying projection. This can be defined arbitrarily, such as by irregular political boundaries. However, the tiles are arranged in a regular tile matrix within each projected planar CRS. This mismatch causes tiles at projection boundaries to extend beyond the actual projection-defined extent. Such protruding tiles are yet crucial for ensuring complete coverage as otherwise gaps between adjacent TGs arise. As a result, parts of the Earth's surface along projection boundaries are also covered by tiles from adjacent TGs. Data ingested into

a TGS, is therefore duplicated due to this redundant coverage. To effectively use the data in these overlapping areas, it is necessary to transform it to a common system and additionally resolve the generated ambiguities. Therefore, while using a higher number of TGs when designing a TGS increases processing complexity and data duplication, it also enhances geometric accuracy.

The Sentinel-2 (S-2) program by the European Space Agency exemplifies a multispectral imaging mission that employs a TGS based on the widely used Universal Transverse Mercator (UTM) projection system and Military Grid Reference System (MGRS) (ESA, 2015). This mapping setup divides the Earth into 60 longitudinal zones of a narrow 6° wide extent. Each zone is further split into a northern and southern part, both realising a TGs. Apart from the resulting overlapping areas along the TGs boundaries, the original MGRS tile matrices are adapted so that adjacent tiles within the same TG overlap as well. This design creates two types of spatial ambiguity, as geographical locations may be realised several times in different or even the same CRS (Roy et al., 2016).

Consequently, retrieving information from overlapping areas becomes increasingly complex, and the overall data volume associated with S-2 imagery grows accordingly. A previous analysis, based on area calculations, examined the amount of the redundantly mapped land surface by the S-2 TGS and identified an overhead of 33% across global landmasses (Bauer-Marschallinger and Falkner, 2023). Relating to the annual volume of user-level S-2 products, which slightly exceed 4 PiB (Castriotta, 2023), this corresponds to more than 1 PiB of redundant data that must be stored, processed, and distributed to users.

Building on that land surface experiment, the present thesis seeks to verify and expand its findings. Rather than estimating the land surface mapped by the S-2 TGS and comparing it to Earth's actual land surface, this work analyses actual S-2 imagery. However, the procedure itself is similar, as the data volumes of various datasets stored within the S-2 TGS are compared against reference volumes in order to quantify the amount of excessive data. The identical experimental setups as in the land surface analysis are repeated here, enabling direct comparisons of both metrics while providing additional insights into the consequences for real imagery.

Chapter 2 starts with detailed definitions of relevant TGSs. The full methodology, including the derivation of reference values for the redundancy estimation, the experiment setups, the data, and all preparation steps, is outlined in Chapter 3. Chapter 4 presents the results of the data volume analyses, including comparisons to the findings of the land surface experiment. This is followed by a concluding discussion in Chapter 4 that highlights the main insights and addresses remaining open questions.

2 Tiling Grid Systems

This chapter explores the concept of TGSs by examining popular examples and outlining their advantages and disadvantages. Starting with the main objective of this thesis, the first system, used for distributing S-2 user-level data, originates from military and navigational purposes. As a result, it features several drawbacks when utilised for Earth Observation imagery. In contrast, the second presented example was specifically developed and optimised for multi-resolution geospatial raster data. Therefore, with slight adaptations it is used as a reference benchmark for the data volume experiments. The chapter concludes with a description of the TGS used for the Landsat data, which is the second leading global land monitoring program alongside S-2 (Claverie et al., 2018). This TGS is not part of the experiments, but its development highlights the significance of efficient spatial frameworks.

2.1 The Sentinel-2 Tiling Grid

2.1.1 Universal Transverse Mercator system

The Sentinel-2 TGS is based on the widely used Military Grid Reference System (MGRS) and its underlying Universal Transverse Mercator (UTM) projection system. This global mapping system divides the Earth into 60 zones, each covering 6° longitude. Within each zone, an individual transverse Mercator projection is applied, which minimises geometric distortion due to the limited narrow extent of each zone, thus resulting in high local geometric accuracy (Snyder, 1987).

In general, a transverse Mercator projection wraps a cylinder around a reference ellipsoid, with its axis of rotation perpendicular to that of the ellipsoid. Ellipsoidal coordinates are projected onto the cylinder, which is then unrolled to yield a planar representation of the ellipsoid in metric units. The resulting map remains undistorted along the reference meridian, where the cylinder tangentially touches the ellipsoid. This preserves angles and shapes to a local extent due to the conformal map characteristics. However, distortion increases with distance from this reference meridian.

To uphold spatial accuracy, the UTM system limits each projection's extent to ±3° of the meridian. Hence, the globe, represented by the World Geodetic System 1984 (WGS84) datum, is split into 60 zones, employing identical projection formulas but different parameters. The system achieves full longitudinal coverage as each zone rotates the cylinder by 6°. Additionally, the cylinder is scaled down by a factor of 0.9996, causing it to intersect the Earth's surface twice rather than touching it along the central meridian. This results in true scale along the secants, smaller scale inside and bigger scale outside, thereby minimising overall distortion within each zone. An illustration of the UTM cylinder setup is presented in Figure 2.1a.

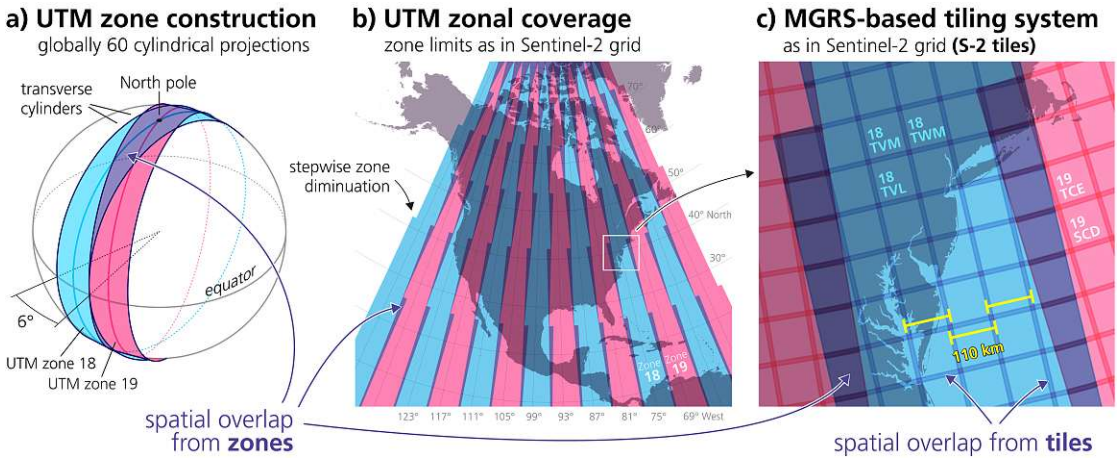

Figure 2.1: Setup of the UTM system and the MGRS based S-2 TGS (taken from Bauer-Marschallinger and Falkner, 2023)

As Figure 2.1c depicts, the meridian strip setup is well suited for areas with a wide-ranging north-south extent. In contrast, areas stretching far east-west are mapped into multiple zones. Due to the nature of cylindrical projections, areas near or at the poles would be split across all zones. Thus, the UTM system limits its latitudinal extent to 84°N and 80°S, with the polar regions being covered by the Universal Polar Stereographic (UPS) projection system, complementing UTM to achieve global coverage.

The coordinate system of each zone is realised by the projected equator as the Easting axis and the reference meridian as the Northing axis, arranged orthogonally to form a Cartesian coordinate system. To avoid negative coordinates west of the central meridian, an offset of 500,000 meters is added to the Easting coordinate (False Easting). Similarly, 10,000,000 meters are added to the Northing coordinate in the southern hemisphere (False Northing). This offset is applied only to coordinates south of the equator. Therefore, it is necessary to unwind the False Northing when using both hemispheres of a UTM zone. Consequently, each zone features two coordinate systems, resulting in a total of 120 projected CRSs.

The UTM zones are labelled from 1 to 60, starting with the central meridian of the first zone at 177°E and moving eastwards. An additional N or S designation indicates the northern or southern hemisphere. Each CRS is assigned a European Petroleum Survey Group (EPSG) code defining the underlying geodetic datum, projection equations, and parameters (IOGP, 2022). The corresponding UTM EPSG codes are formulated as 32000 + 600 (or +700) for the northern (southern) hemisphere + the zone number. For example, EPSG code 32633 corresponds to UTM zone 33N, which covers longitudes from 12° to 18°E and latitudes from 0° to 84°N.

Although initially developed for military purposes, the UTM system has become a quasi-standard across various applications, facilitating the joint use with remote sensing imagery (Buchroithner and Pfahlbusch, 2017).

2.1.2 Military Grid Reference System

The Military Grid Reference System (MGRS) further refines the UTM system and is employed by NATO military forces for precise point location referencing. A detailed and complete definition of the MGRS and its characteristics can be found in (DMA, 1989).

First, the 6° wide UTM zones are latitudinally split into grid zones of 8° height. These latitude bands are labelled from C to X, with A, B, Y, and Z reserved for the UPS extension. Starting at the southern zone borders at 80°S and moving northwards, the equator serves as an exact separator between the M and N bands. The northernmost band, X, extends to 12° at its northern borders of 84°N. Combined with a zone number, a latitude band forms a grid zone designator covering a geographic area of 6° × 8°.

Second, each projected UTM zone is further divided into squares of 100 × 100 km, always with corner coordinates as multiples of 100,000 meters. These squares are labelled with two additional letters, defined by a zone-dependent alternating scheme, representing the row and column of a square. Thus, a five-digit alphanumeric code consisting of the zone number, the grid zone designator, and the row and column letters identifies each square. While the MGRS provides further referencing of point locations within each square with a precision of up to one meter, this level of detail is not relevant to this thesis and is therefore omitted.

2.1.3 Sentinel-2 tiles

The S-2 mission is designed to provide medium-resolution optical imagery for land monitoring. It delivers essential data for environmental management, agriculture, forestry, and disaster response. The mission includes two identical satellites, Sentinel-2A and Sentinel-2B, with two additional satellites planned in the future. This configuration ensures a revisit frequency of 5 days at the equator. Each satellite is equipped with a multispectral instrument. These instruments capture data across 13 spectral bands ranging from visible to shortwave infrared wavelengths with a swath width of 290km. According to the band, the spatial resolution of the delivered data is 10, 20, or 60 meters (ESA, 2015).

For distributing S-2 user-level data, a TGS based on the MGRS 100km squares is used. Its definition is available for download from the S-2 products page (ESA, online) and includes all 56,686 tiles as polygons. Each S-2 tile is an enlarged version of the corresponding MGRS tile, extending by 9.8km to the south and east, while the northern and western borders

remain unchanged. This enlargement effectively shifts the tile centroids southeast, resulting in tiles that measure 109.8 × 109.8km with a 9.8km overlap between neighbouring tiles, as illustrated in Figure 2.1c. While this information is easily obtainable by examining the official file and actual S-2 imagery, the product specification documentation states 100 × 100km tiles (TAS, 2022). Meanwhile, the S-2 products page introduces a further contradiction by mentioning 110 × 110km tiles.

As discussed in Bauer-Marschallinger and Falkner (2023), the S-2 documentation does not provide a rationale for the tile enlargement or its specific magnitude. Accordingly, it can only be assumed that this is intended to correct for shadows cast by clouds located near the original south and east tile borders, as shadows are cast westward at the S-2 mean local solar time of 10:30 am. Presumably, an extension of 10km was aimed for, resulting in tiles of 110km extent. However, specifying the tile extent indirectly purports the sampling of the underlying PG. In this case, 110km tiles are non-consistent with a PG of 60m sampling ($110,000 \bmod 60 \neq 0$). Therefore, the seemingly uncommon 9.8km enlargement can be explained as the next smallest common multiple of 110km to the 10, 20 and 60m spatial resolutions.

However, this overlapping scheme generates multiple realisations of the same cells in the underlying PG of each UTM zone, resulting in data duplication. After ingestion into the TGS, S-2 products are processed at the tile level. Enabling this is one of the goals of partitioning data into tiles, but the multiple processing of the same pixel value may yield unequal results. PG cells occurring in several tiles may then hold different values, although deviated from the identical observation. Solving this ambiguity for processing or displaying data requires aggregating all existing pixel values into a single one.

The underlying UTM setup introduces the other type of spatial ambiguity within the S-2 TGS. The division of the Earth's surface into a total of 120 zones and thus an equal number of TGs results in a large number of TG boundaries. Along these, tiles belonging to different TGs overlap. Since the width of UTM zones decreases with increasing latitude, the number of tiles per row in the tiling matrix shrinks aswell. While this counteracts unnecessary large area overlaps, it is again reinforced by the tile enlargement of the original MGRS tiles, as border tiles protrude even further into adjacent TGs. Overall, a considerable amount of overlap arises, as shown in Figures 2.1b/c. Opposite to the intra-TG overlap, geographic locations are represented in different PGs. As these are defined in different spatial domains, resolving demands appropriate reprojection and resampling methods before aggregation to preserve geometric fidelity (Roy et al., 2016).

Figure 2.2 illustrates the combined pattern of both types of overlap. As shown, single locations can be covered by up to six tiles. This manifold realisation of single geographic locations is detrimental because the S-2 processing maps each observation onto all corresponding tiles

(ESA, 2023), leading to two significant issues. Firstly, observations are stored and processed multiple times, resulting in a substantial data duplication rate of 33% over global landmasses (Bauer-Marschallinger and Falkner, 2023). Secondly, data handling becomes more complex as users encounter multiple pixel values for the same location, all representing the same observation. Failure to address this issue may lead to inaccurate analysis or processing results.

2.2 The Equi7Grid

The E7G is a spatial reference framework optimised for high-resolution remote sensing raster images. It is designed to efficiently handle archiving, processing, and displaying raster data while preserving geometric accuracy with a minimal data oversampling of just 2%. A detailed analysis of the advantages of the E7G compared to other possible partitioning and projection options is provided in (Bauer-Marschallinger et al., 2014), while technical specifications are documented in (TUW, GitHub, online).

2.2.1 Continental partitioning

The E7G employs a global partitioning approach that organises the Earth's surface into seven continental-based zones. Each zone defines the area to be mapped by a cartographic projection. The boundaries are deliberately drawn over oceans where possible to maintain the integrity of landmasses. Furthermore, political entities as countries are not split into multiple zones whenever feasible. Once defined, the boundaries are buffered by 50km over landmasses to facilitate accurate spatial filtering operations. An illustration of this resulting continental partition is presented in Figure 2.3.

Driven by practical utility, this approach prevents splitting common areas of interest, such as countries or entire continents, and thus simplifies the overall data handling. A purely geometric partitioning of the Earth's surface, as practised by the UTM setup, does not offer this advantage. However, manual delineation of boundaries may give rise to geopolitical disputes regarding the exact borderlines along ambiguous natural borders, such as the boundary between Europe and Asia.

Using a quantity of seven zones for global coverage strikes a balance between data handling complexity and geometric distortion. A continental zone covers a significantly larger geographic area than local partitions such as UTM zones. This potentially results in more pronounced geometric distortions when applying projection functions, but reduces complexity in data handling. A global projection featuring only a single zone would minimise complexity while yielding tremendous distortion. Thus, seven continental zones represent a practically orientated trade-off between global and local partitioning. Furthermore, limiting the number of partitions reduces the number of projection boundaries. As demonstrated by

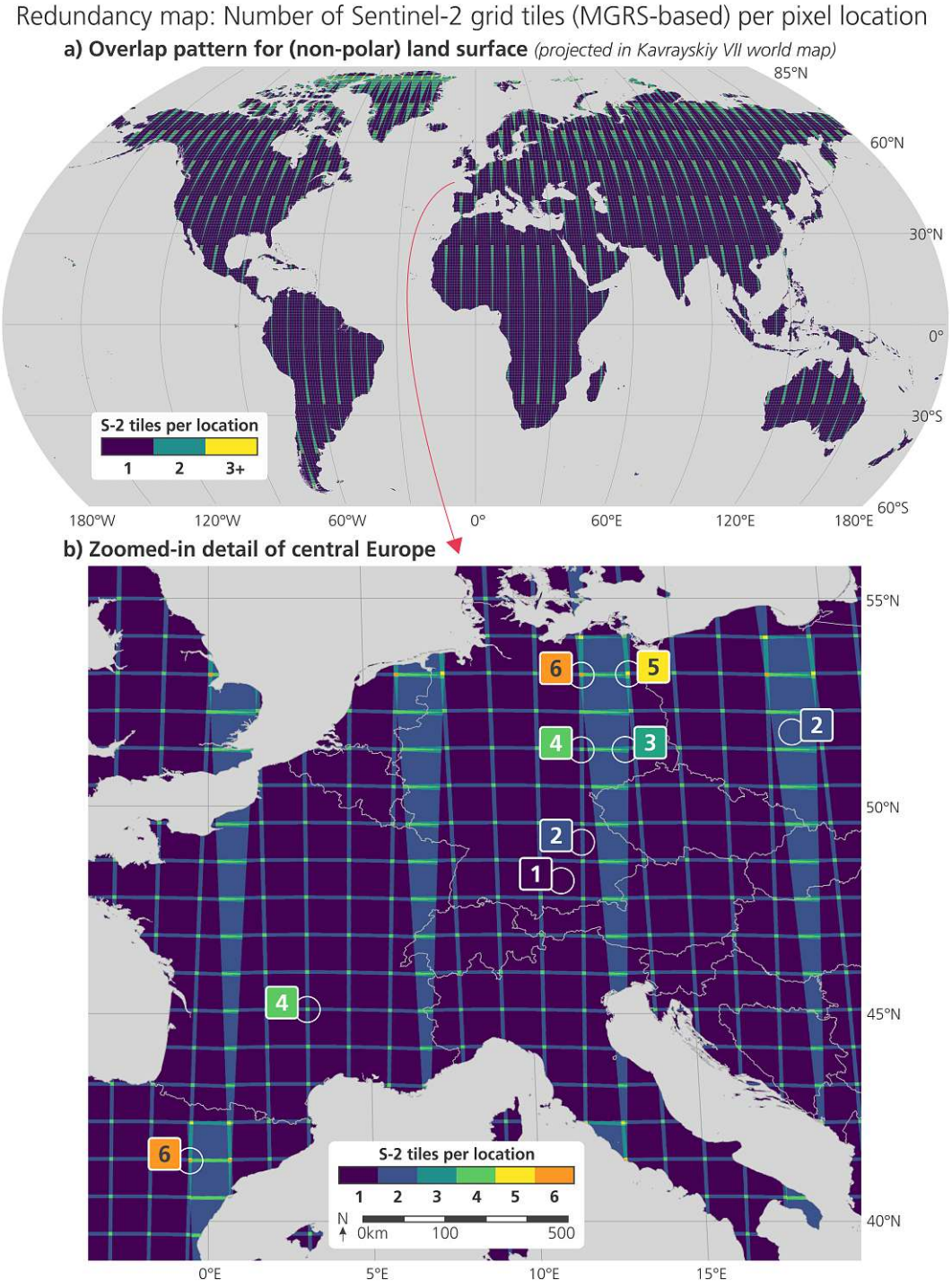

Figure 2.2: Number of overlapping S-2 tiles per location (taken from Bauer-Marschallinger and Falkner, 2023)

the UTM setup, this is particularly relevant over landmasses, as it decreases data duplication and lessens the need for data aggregation in these areas.

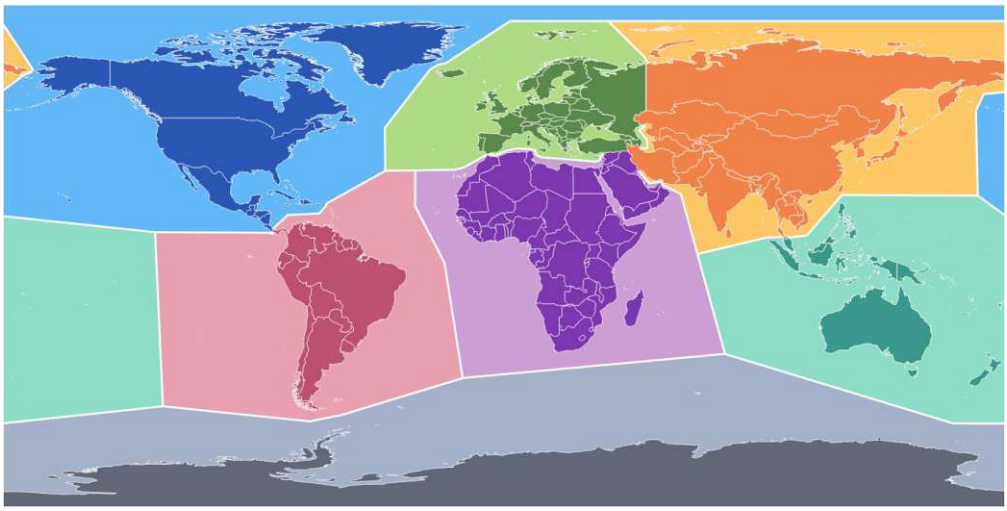

Figure 2.3: Figure 3 from (TUW, GitHub, online) visualising the continental zone partitioning of the Equi7Grid in the geographic latitude-longitude domain.

2.2.2 Equidistant Azimuthal projection

The Equidistant Azimuthal projection proves most suitable for minimising data oversampling for regular PGs across continental partitions of the globe. In general, an Equidistant Azimuthal projection maps ellipsoidal coordinates radially onto a plane tangent to the reference ellipsoid at a chosen central point. Locations along an azimuth are projected onto a straight line from this chosen projection centre, while their proportionally correct distances to it are preserved. However, with increasing distance, the distortion of angles, shapes and areas of mapped objects becomes more pronounced (Snyder, 1987).

Yielding high spatial accuracy, the E7G selects an individual central point for each continent tailored to the geographic characteristics of its landmasses. This results in optimised projection parameters, minimising each PG's oversampling. Coinciding with the UTM system, the WGS84 datum is used as an ellipsoidal representation of the Earth. Applying the projections to the continental zone results in seven planar CRSs in metric units. The Northing axis coincides with the meridian passing through the central point, while the Easting axis is orthogonally aligned. False Easting and False Northing offsets are applied to the projections in order to maintain all coordinates within the positive domain. However, since the zones are different in shape and size, individual values are used for each.

Figure 2.4 depicts the mapping process of the European continental zone from the geographic to the projected domain. It is to be noted that the centre of the projection is not located at the zones, but the landmasses centroid in order to minimise distortion over the latter. Furthermore, the Mediterranean forms a natural border towards the African zone, not necessitating the splitting of countries. However, the situation differs with respect to the Asian zone, as the extent of Russia clearly spans two continents, as per E7G definition.

The exact projection definitions can be accessed with the EPSG codes 27701 to 27707. In addition to the datum definition, these include the projection's central points and applied axis offsets. Along with the precise zone boundaries, the codes encompass all necessities for implementing continental PGs, to which tiling schemes can be applied subsequently.

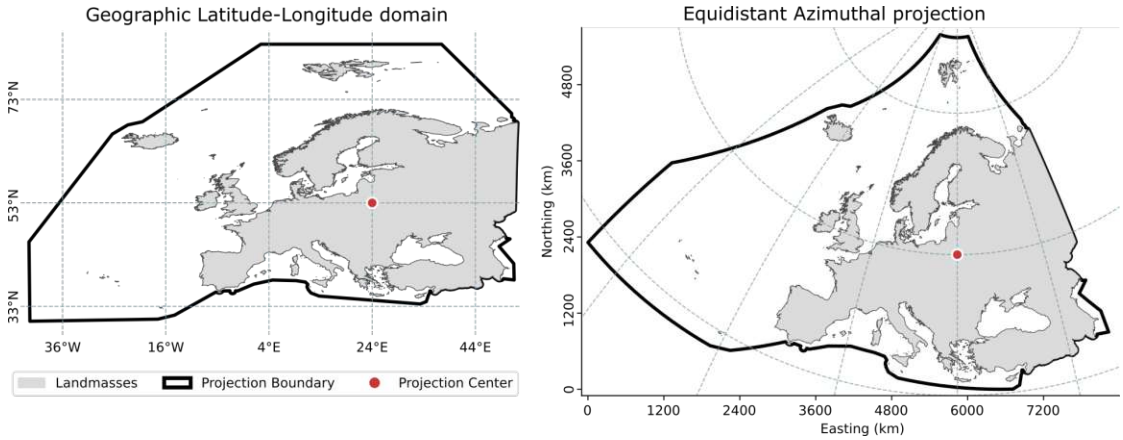

Figure 2.4: European zone of the E7G in the geographic latitude-longitude and projected domain.

2.2.3 Tiling levels

Unlike the S-2 TGS, which primarily adapts its tile sizes to the characteristics of the S-2 mission products, the E7G is designed for more generic Earth Observation raster data. Recognising the variation in spatial resolution among remote sensing imagery, three tiling levels are introduced. Each level features a tile matrix consisting of axis-parallel squares with extents of 100 km (T1), 300 km (T3), or 600 km (T6). With the requirement in mind that the tile size must be an exact multiple of the pixel sampling, this range of option enables a wide variety of imagery to be used within the E7G. As part of future updates to the system, even more flexible tile sizes and consequently sampling sizes are scheduled.

The most appropriate tiling level for the imagery is chosen based on its spatial resolution. For instance, T1 tiles are most suitable for 10m sampled data, while T6 tiles are more appropriate for coarser samplings such as 500 meters. By varying the tile size dependent on the spatial resolution, the number of pixels to be handled at once always remains within reasonable limits. However, no overhead is introduced by over-splitting the underlying PG into unnecessarily small tiles, either.

Strictly speaking, the E7G introduces three TGs per zone instead of one. These are nonetheless congruent and hierarchically linked. Specifically, a 3×3 grid of T1 tiles forms a single T3 tile, while a 2×2 grid of T3 tiles forms a single T6 tile. Regardless of the tiling level, all tiles are aligned such that their corner coordinates are multiples of the tile size. This design enables a straightforward naming convention based on the bottom-left corner coordinates,

allowing for the intuitive identification of tiles. For example, E054N015T3 denotes a tile with its bottom-left corner at 54,000km Easting and 15,000km Northing, with a tile size of 300km (T3). To ensure global uniqueness, a zonal prefix (e.g., EU for Europe or AF for Africa) must be additionally added, as all continental zones use the corner-based nomenclature. Exemplarily, figure 2.5 continues figure 2.4 by displaying the tiling levels in addition to the projection.

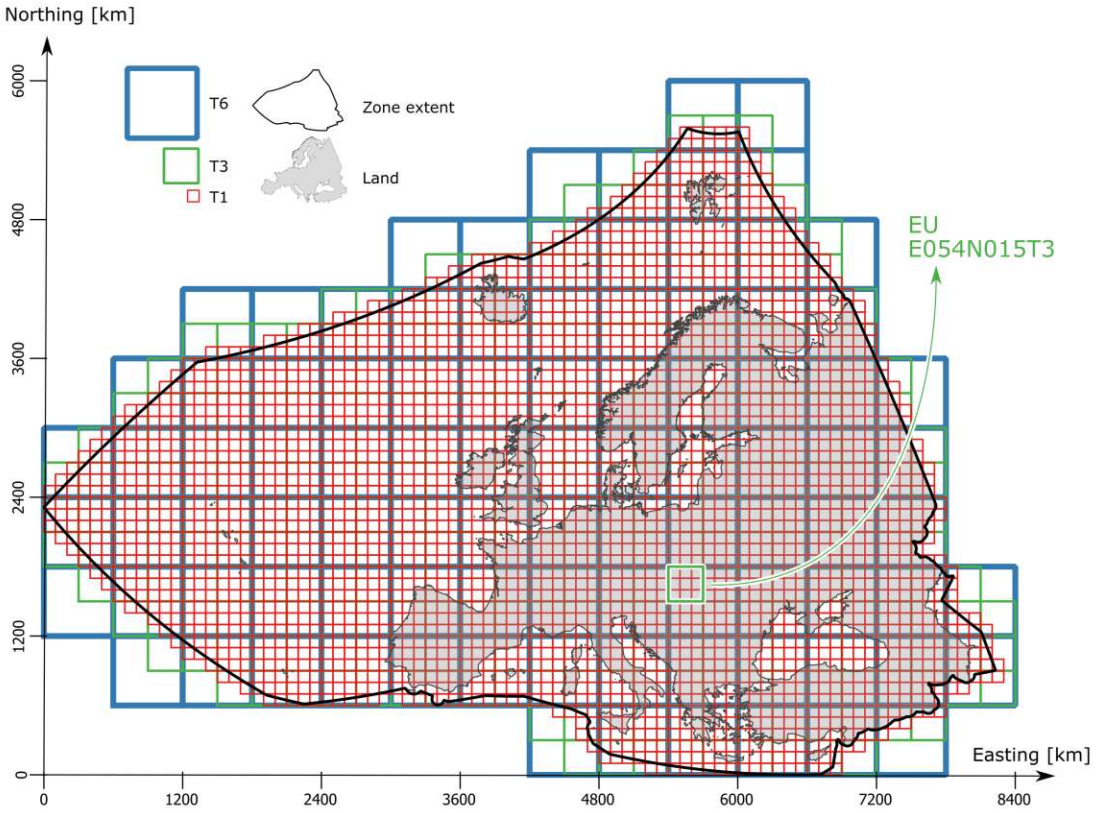

Figure 2.5: Equi7Grid tiling levels of the European zone in the projected domain.

Figure 2.5 also indicates that tiles of the tile matrix that do not intersect the projected zone extent are not generated, as their coverage is fully realised within adjacent zones. However, as already mentioned regarding the MGRS setup, partially intersecting tiles are essential to guarantee full zone coverage and achieve a complete representation of the Earth's surface in the TGS. Notably, the areas that are mapped by multiple TGs due to these partly extending tiles are significantly smaller compared to the S-2 TGS. This reduction is attributed to the fact that zone boundaries avoid crossing landmasses wherever possible. As shown in Figure 2.3, there is a maximum overlap of three TGs along the Iranian-Iraqi-Turkish border, where the African, Asian and European zones meet.

Since tiles within the E7G are designed non-overlapping and only seven TGs are required to reach global coverage, the system yields with 3% significantly less redundantly mapped land

surface, with only 3% compared to 34% in the S-2 TGS (Bauer-Marschallinger and Falkner, 2023). Additionally, data handling is simplified as areas of interest are more likely to be fully covered by a single TG, making this TGS an efficient choice as a spatial reference framework for remote sensing data.

2.3 Landsat ARD Tile Grids

With over five decades of operation, the Landsat program is the longest continuous monitoring mission of the Earth's landmasses. By providing optical imagery, it supports a wide range of applications, including agriculture, forestry, water resources management, and disaster response. The latest satellite in operation, Landsat 9, collects data across 11 spectral bands. These include visible, near-infrared, shortwave infrared, and thermal wavelengths. With a 16-day ground track repeat cycle and a swath width of 185 km, the mission delivers imagery at a spatial sampling of 30 meters. Combined with Landsat 8, which carries nearly identical sensors, the revisit time is reduced to 8 days (Masek et al., 2020).

2.3.1 World Reference System 2

The World Reference System 2 (WRS-2) establishes the dissemination of Landsat data by systematically organising imagery into a structured grid system, defined by paths and rows, corresponding to the satellite's orbit. Paths represent orbital tracks, while rows denote latitude-specific positions (NASA GSFCb, online). This structure allows the segmentation of continuously acquired sensor data into distinct scenes as the satellites move along their orbit. Each scene is uniquely identified by its path and row number in conjunction with the acquisition date. Along with radiometric and geometric corrections, scenes are registered to the UTM projection system and made available to users (Sayler and Glynn, 2022).

However, the WRS-2 features one significant drawback: the temporal inconsistency of scene footprints due to orbital perturbations that affect the satellite's position and sensor orientation. Consequently, pixels from scenes with identical path/row coordinates captured during different overpasses may not align spatially. Alignment procedures enabling image stacking to create datacubes and conduct multi-temporal analyses thus remain the responsibility of the users (Wulder et al., 2022).

2.3.2 ARD Tile Grids

Recognising the need for easily usable Landsat data, an additional product family meeting the criteria of ARD was introduced in 2017. These so-called Landsat ARD tiles are preprocessed with atmospheric corrections, include quality indicators, and utilise a TGS rather than the WRS-2. The structure of this framework is detailed in (Dwyer et al., 2018).

Unlike the previously discussed TGSs, the Landsat tiles do not provide global coverage. Currently, their extent is limited to the United States of America, including the contiguous United States (CONUS), Hawaii, and Alaska, with plans for global expansion in progress (Zhu, 2019). Each of these areas forms an independent zone, mapped separately due to the geographic isolation of their landmasses. While this segmentation increases complexity, it allows the use of optimised projection parameters for smaller projected areas, reducing geometric distortions.

The Albers Equal Area Conic projection was selected for mapping. This projection transfers ellipsoidal coordinates onto a conical surface intersecting the reference ellipsoid along two standard parallels. Meridians are projected as equally spaced radii, and parallels form concentric arcs when the cone is unrolled to realise the map plane. Although the scale along parallels is too small within the standard parallels and too large outside them, the opposite applies to meridians. For this reason, the projection yields an equal-area map, ideal for area-related analyses. However, distortions in shape and distance increase with distance from the standard parallels. Thus, the Albers Equal Area projection is most suitable for regions extending mainly east-west along the defined standard parallels (Snyder, 1987).

The underlying CRS for each zone is realised by applying the projection with individually selected parameters concerning the WGS84 datum. Since neither False Easting nor False Northing offsets are applied, negative coordinates must be dealt with. The non-overlapping tiling schemes feature tiles that measure 150km, corresponding to 5,000 by 5,000 pixels at Landsat's 30m resolution. Tiles are labelled with horizontal (h) and vertical (v) indices based on their position within the tile matrix for CONUS, Hawaii and Alaska, respectively. The horizontal index progresses eastward from 0, while the vertical index increases southward from 0.

Because integer degrees are used as projection centres and no coordinate offsets are applied, tile corners and pixel coordinates feature non-round numbers. For example, the upper-left corner of Alaska's northernmost and westernmost tile (0h, 0v) is located at -851,715m Easting and 2,474,325m Northing. The TG definitions are available for download as shape files (USGS, online). Figure 2.6 illustrates the tiles alongside landmasses in their respective projected domains.

The non-contiguous nature of the zones prevents their TGs from intersecting. Additionally, there is no tile overlap within the TGs. This ensures that no areas are covered by multiple tiles, neither in the same nor in different projections. Therefore, no redundant data due to zonal partitioning and tiling is produced when ingesting data into the TGS. However, it is important to note that equal-area projections oversample raster data to a greater extent compared to conformal or equidistant projections. In addition, the failure to preserve angles during mapping skews neighbourhood relationships, impacting the application of local

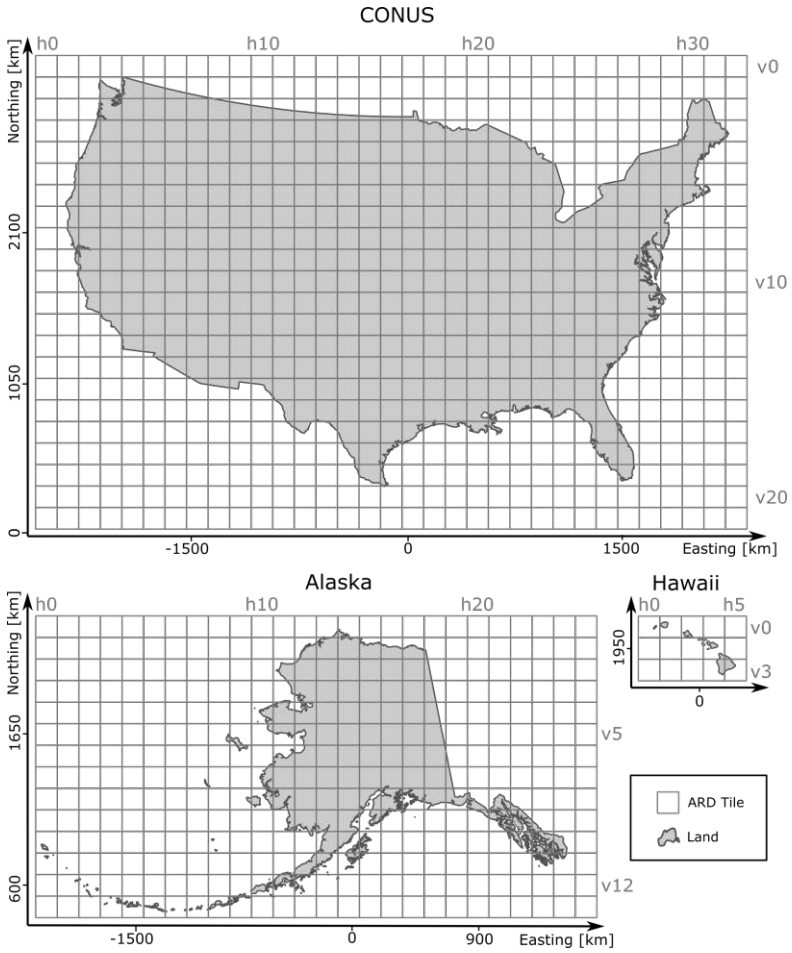

Figure 2.6: Tiles defined for the Landsat ARD data

filters and operators (Bauer-Marschallinger et al., 2014). Therefore, utilising equal-area projections as the foundation for the TGS underlying regular pixel grids appears to be a suboptimal choice.

3 Methods and data

3.1 Experiments based on land surface data

To quantify the amount of redundant data produced by the S-2 mission due to its TGS, Bauer-Marschallinger and Falkner, 2023 measured the global land surface mapped by the system. By comparing the result to the Earth's actual land surface, the system's spatial overhead was estimated. Since the mapped area is directly proportional to the data volume in terms of regular PGs (Kimerling, 2002; Mulcahy, 2000), this metric provides insights into the TGSs redundancy, highlighting the amount of data that is redundantly generated.

The starting point of the spatial overhead experiments was the creation of a global binary raster differentiating between land and non-land surfaces. For this purpose, the Natural Earth land dataset detailed in Chapter 3.3.2 was chosen. Provided in the geographic latitude/longitude domain (EPSG:4326), it was reprojected to the Plate Carrée projection (EPSG:4087) and rasterised with a sampling of 1km.

While reprojecting this base raster to each UTM zone, it was clipped to a generously buffered ±15° extent around each reference meridian. This ensured complete coverage of all S-2 tiles extending beyond their zone, which would not have been achieved with the original ±3° stripes. Subsequently, summing up the amount of land pixels within each tile across all zones revealed the total land surface mapped by the TGS, as pixels covered by multiple tiles were also counted the respective number of times.

The previously generated raster in Plate Carrée also facilitates estimating the Earth's actual land surface, representing the 100% baseline for the overhead assessment. For this purpose, each land pixel is scaled by the cosine of its latitude, considering the projection's non-equal-area characteristic, as it maps parallels as equidistant straight lines (Snyder, 1987). Given the 1km sampling, each pixel corresponds to the distorted area of 1km^2. Therefore, the total of all scaled pixel values provides the sought reference value.

Exploring the TGSs redundancy, the experiment was conducted with modified setups. First, the base raster was clipped to include only latitudes between 60°S and 85°N. This limitation omits polar regions, which are not part of the S-2 coverage (ESA, 2015) but includes the remainder of the Earth's major landmasses. The mapped land surface and reference value for this subset were then recalculated identically.

A second experiment analysed the spatial overhead on a continental level. For each of the six remaining continents, the limited base grid from the first setup was masked to exclude land pixels corresponding to other continents. These adjusted rasters were then used as input to determine the mapped land surfaces and the corresponding reference values for each continent independently.

Finally, a third approach aimed to differentiate between the impacts of overlapping tiles and overlapping TGs. For this purpose, all tiles of a TG were dissolved into a single polygon, eliminating the overlap within the TG. Repeating the calculations with this simplification thus yields values corresponding to the spatial overhead caused only by tiles overlapping with tiles from different TGs. Since both types of overlap add up to the previously determined total overhead, comparison to the results of the first setup as well reveals the magnitude of the tile overlap within the TGs.

3.2 Experiments based on image data

The experiments described in Section 3.1 estimate the redundancy within the S-2 TGS over landmasses by using the mapped land surface. However, in reality, not the pure amount of pixels, but the data volume is the decisive factor for storing data. Thus, an indirect unit of measure was used, and factors applying to real imagery, such as file format and image compression, were not considered in the experiments. Furthermore, a spatial sampling of 1 km was used, which is also not in line with the spatial resolution of Earth Observation products that require the usage of TGSs. Hence, addressing both points of simplification, the experiments conducted in this thesis estimate the system's redundancy by evaluating the data volume of actual S-2 datasets.

An intrinsic overview about the decision on selection and characteristics of the used imagery is given in Chapter 3.3.1, while necessary preparation steps as well as exact format specifications are reasoned in Chapter 3.4. In brief, each examined dataset is a single-layer raster with global coverage of the Earth's land surfaces, tiled according to the S-2 TGS. This means for each at least partially land-covering S-2 tile, a single image file exists, which assigns representative pixel values to land areas. Non-land areas within these tile coverages are masked, as they are inevitably included due to the earth-fixed tiles, but are not of interest in experiments related to land areas.

The methodological framework used to estimate the redundant data volume closely follows the approach taken in the land surface experiments to estimate the so-called spatial overhead. However, instead of counting land pixels, here the total data volume across all tiles in a dataset is summed. This metric reflects the actual storage requirement when data is maintained within the S-2 TGS. To replicate real-world conditions, format specifications commonly applied in operational settings are used. In order to quantify the proportion of data that is unnecessarily generated by the system, the data volume is compared to a reference value, analogous to the use of the Earth's actual land surface. This yields the so-called data volume overhead, which conceptually corresponds to the spatial overhead metric.

3.2.1 Estimation of reference values

The reference value required for the overhead estimation represents the theoretical volume required to store a dataset within a spatial framework that is entirely free of redundancy. However, as outlined in the introduction, the efficient use of a global raster dataset at a resolution comparable to that of S-2 fundamentally necessitates the application of a TGS as a spatial reference framework. TGSs, in turn, are based on map projections that represent the Earth's surface as planar, two-dimensional regular grids. This unavoidably introduces geometric distortions, whose magnitudes depend on the type and coverage of the chosen projection. When applied to regular PGs, such distortions manifest in redundancy. Nonetheless, a TGS can be designed as a composite of optimised TGs, each tailored to a specific region of the Earth's surface. While this approach minimises the TGSs projection-driven distortion, it does not eliminate it entirely. As a result, a completely redundancy-free TGS and truly redundancy-free reference value cannot exist.

Assuming that the projection-driven distortion of the S-2 TGS tends towards zero due to the many and narrow UTM zones, only the overlaps between tiles of the same and different TGs remain as sources of redundancy. In theory, these redundancies could be removed by creating a zonal mosaic per TG, then clipping the results to the precise boundaries of the corresponding UTM zones at a pixel level. This two-step approach would ensure that every part of the Earth's land surface is mapped exactly once, as the UTM boundaries represent a globally gapless and unambiguous spatial partitioning. However, the pixel clipping poses a problem, depicted in 3.1. Namely, the S-2 adaptation of the MGRS defines certain tiles that lie entirely outside of the nominal boundaries of their corresponding 6° wide UTM zones. Such tiles are then fully located in the coverage of an adjacent UTM zone, although not associated with the adjacent TG. In such cases, the S-2 TGS omits the definition of native tiles if these are already fully covered by out-of-zone tiles, avoiding full double coverages. However, because of these adjustments, not all zonal mosaics achieve gapless coverage of their UTM zones, although sometimes extend into adjacent zones and vice versa. Thus, limiting the mosaics to the UTM boundaries removes the redundant mapping of land surfaces to tiles of different TGs, but also yields gaps in the global coverage. Hence, the precise UTM zones are not suitable to remove the overlap of tiles stemming from different S-2 TGs. However, due to the irregularly added and omitted tiles, it is also not possible to adapt the boundaries accordingly, impeding the estimation of reference values from data stored in the S-2 TGS.

E7G as reference system

Consequently, the E7G is adopted as the spatial framework used for the reference determination. It is specifically optimised for land surfaces, which aligns directly with the defined object of interest in this work. As a result, it can be assumed that the E7G introduces only minimal redundancy over those areas. Supporting this assumption, the system

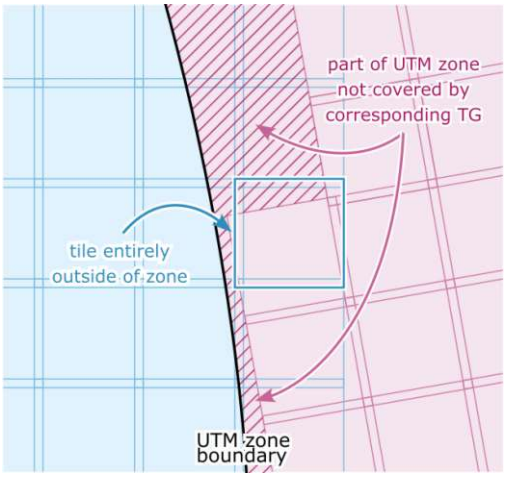

Figure 3.1: Schematic depiction of the unsuitability of the S-2 TGS to remove zonal overlaps.

has already been the subject of multiple studies assessing its redundancy characteristics. These investigations yielded consistently low estimates. Namely, a mean oversampling of 2% (Bauer-Marschallinger et al., 2014) and a surface overhead of 3% (Bauer-Marschallinger and Falkner, 2023) were estimated, indicating that the E7G is close to the optimum of being free from redundancy. Thus, substantiating its choice as a suitable reference framework for the analyses conducted here.

Admittedly, the reported redundancy figures are not solely attributable to the use of the equidistant projections, but also arise from overlapping tiles along the boundaries of the TGs. In contrast to the S-2 TGS, the E7G enables the removal of the latter by clipping tiles to the continental boundaries at the pixel level, as the system defines only those tiles per zone that are strictly necessary to achieve complete coverage. Nevertheless, a design convenience still hinders the full removal of the overlap redundancy as depicted in Figure 3.2. The originally defined global partitions, which are intended to uniquely define the extent of each TG, were buffered by 50km over landmasses. As a consequence, the actual spatial extents of the TGs and corresponding tiles overlap with one another by these buffers. This design choice facilitates processing in border regions by ensuring data availability slightly beyond the nominal boundaries, but comes at the cost of introducing otherwise avoidable redundancy. Yet, it is possible to simply use the original, unbuffered zone boundaries as clipping geometries to trim the corresponding tiles, thereby completely removing the overlaps and their evoked redundancies.

However, before applying any adjustments practically, all datasets must be converted to the E7G. To ensure a fair comparison using similarly sized tiles, the E7 T1 tiling level is employed. The specific details of the applied conversion workflow are described in Chapter 3.5. The conversion results are then versions of the datasets representing the standard usage of the

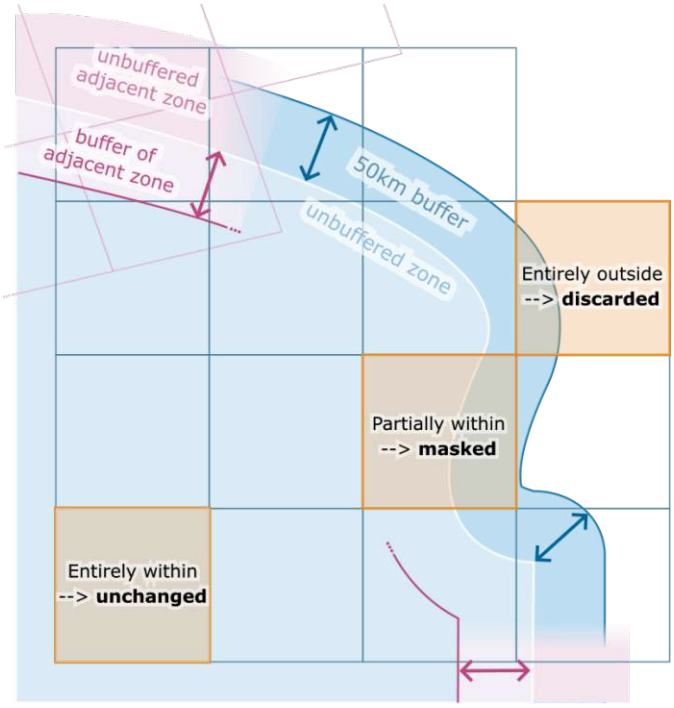

Figure 3.2: Schematic depiction of the Equi7Grid with examples of tiles to be unchanged, masked and discarded to remove redundant mapping.

E7G and therefore are referred to as the *E7 versions*, in contrast to the conversion inputs, which are labelled *S-2 versions*, since stored in the S-2 TGS.

Clipping to unbuffered partioning

Once the E7 versions are on hand, the clipping operations can be performed, using the originally defined, unbuffered global partitions as strict boundaries for each TG. Tiles that lie entirely within their respective TGs are retained without modification, as they do not contribute to redundancy. Conversely, tiles located entirely outside their TG's original extent are discarded, as these only map the redundant buffer zones. Tiles partly intersecting the unbuffered boundary remain, as they are needed to provide complete coverage. However, they are masked so that no valid pixel values exist outside the TG's unbuffered extent. This ensures that each land surface location is mapped only once and only by a single TG, entirely eliminating any form of spatial overlap. This refined version of the datasets is referred to as the *reference version* and serves as the basis for deriving the data volume reference values.

Compression

While pixels in masked areas remain technically present in the files, they are assigned a consistent no-data value. This allows the minimisation of the required data volume for such regions to a negligible amount. Guaranteeing this minimisation and ensuring that no extraneous data is stored, the files of the reference version are stored with a so-called sparse option. In general, each image file is internally divided into blocks, which are individually

compressed and stored. The sparse specification enables that blocks composed entirely of masked pixels are not written to disk at all, thereby avoiding physical storage allocation. Blocks containing a mix of valid and masked pixels are compressed and stored conventionally. As most geospatial software does not natively support sparsely stored imagery, this format is used exclusively for the files of the reference version, where estimating a benchmark is the objective. The S-2 and E7 versions, which are intended to reflect typical datasets in operational use, do not use the sparse format and are stored using conventional settings. Aside from the sparse-specific configuration of the reference version, all dataset variants adhere to the identical file format specifications outlined in Chapter 3.4.2, ensuring that no bias is introduced due to differing data structure or compression settings.

Remaining projection-induced oversampling

It is important to emphasise that even after the adaptations, the reference version still contains the projection-induced oversampling of the E7G. However, the previously described efforts to quantify this value refer to the buffered continental zones. Therefore, the actual oversampling in the unbuffered dataset versions can be assumed to be even lower, particularly because distortions in equidistant azimuthal projections increase with distance from the projection centre, thus the exclusion of peripheral areas reduces this effect. It must be acknowledged, however, that all determined reference values inherently contain a minimal degree of redundancy.

3.2.2 Experiment setups

Identical to the land surface experiments, the analyses in this study are carried out across multiple setups. For each of these setups and datasets, the total data volume of the S-2 version as well as the corresponding reference version is determined. The obtained reference volumes thereby always declared 100% data volume. Consequently, the comparison reveals the degree of redundancy introduced by the S-2 TGS for each dataset.

Global, non-polar domain

The first setup considers the entire global land surface between latitudes 60°S and 85°N. Since the S-2 mission does not provide consistent coverage beyond this extent, experiments outside of this latitudinal range are neither feasible nor conducted. This limitation, however, enables a direct comparison with the spatial overhead results. The methodology involves identifying all tiles intersecting the defined area of interest and summing up their respective data volumes for both the S-2 and reference versions. The comparison of the resulting data volumes then quantifies the incurred overhead when storing global, non-polar datasets in the S-2 TGS.

Continental domains

The second setup evaluates the data inflation on a continental level, again excluding the polar

domain. Given that the S-2 tiling configuration tends to exhibit more pronounced overlaps at higher latitudes due to the nature of the UTM system (Figure 2.2), this setup aims to examine how this affects different geographic regions. For the sake of consistency, the same continental partitions as in the spatial overhead experiments are utilised. It should be noted that these do not align with the definitions of the E7G, as they follow a conventional division of continents rather than one optimised for land surface mapping. The analysis procedure mirrors that of the first setup, with the difference that each continent is treated as a distinct area of interest. This results in six data volume overheads per dataset, excluding Antarctica due to its absence from S-2 coverage.

Tile vs. zone overlap

Lastly, a third setup aims to quantify the individual contributions of the two spatial overlap types to the total redundancy estimated in the first setup. For each dataset in its S-2 version, all tiles identified previously for the first setup are mosaicked into zonal files for each TG, thereby removing intra-TG overlaps. By summing the data volumes of the zonal files and comparing them to the reference values obtained in the first setup, the portion of volume overhead attributable to inter-TG overlaps can be estimated. Since the total volume overhead measured in the first setup results from both overlap types combined, the intra-TG component can be inferred as well.

3.3 Data

A fundamental prerequisite for conducting the data volume overhead experiments obviously is image data stored within the S-2 TGS, as this system is the subject of investigation. Naturally, one would consider using S-2 user-level products, since their dissemination represents the core rationale behind the TGS's existence. However, in order to assess the redundancy of the system across landmasses, the utilised data must provide complete coverage of the area of interest, spanning from 60°S to 85°N, and must do so with exactly one image per tile. Furthermore, to evaluate how different data characteristics influence the volume overhead, the analysis must include not only a single dataset but multiple ones with varying properties.

However, both provided S-2 user-level products, the top-of-atmosphere Level-1C and the surface reflectance (SR) Level-2A (L2A) granules contain data from single swaths only. Consequently, tiles located at the borders of swaths are only partially covered by valid data (ESA, 2015). Additionally, cloud coverage affects both products, altogether making them unfit for this experiment. Accordingly, S-2 datasets aggregated over time and orbits per tile are needed.

3.3.1 S2MOSAIC L2A 2017

In accordance with the requirements, the S2MOSAIC L2A 2017 collection (S2M) was selected. It aims to generate a global, cloud-free S-2 composite over land areas from January 2016 to September 2017, by utilising an optimal minimum subset of L2A granules (Kempeneers and Soille, 2017). Each of these granules provides atmospherically corrected SR for the extent of a single tile. In addition to the image data at the band-dependent spatial resolutions, each granule includes metadata, auxiliary data, additional outputs and a red-green-blue (RGB) quick look. The additional outputs include a so-called Scene Classification Layer (SCL), which is a byproduct of the L2A processing. This SCL allows to distinguish between clouds, cloud shadows, vegetation, water and snow (ESA, 2015).

During the specified time frame, the sensor aboard S-2A acquired 2,128,556 L2A granules over 29,782 unique tiles. The initial step in generating the S2M involves identifying the highest quality granules per tile. This filtering reduces the overall volume of data to be processed and removes acquisitions without added value. Firstly, granules with a sun zenith angle greater than 70° were excluded as these most likely feature shadow-impacted observations. The remainder were ranked based on their average cloud-free blue band pixel values, serving as an indicator of atmospheric influence and image quality. Since granules regarding the same tiles acquired from different orbits feature different tile coverages, this ranking was performed independently for each relative orbit. Based on the blue band values, only the lower 50% of granules were retained for further processing. However, if insufficient granules remained to create a cloud-free composite for a tile, additional ones were incrementally included from the initially rejected set. After these quality checks, 1,071,123 granules were retained from the entire record.

An optimisation algorithm was then employed to select the minimum set of granules required to achieve full coverage of each tile's extent without cloud contamination. Accordingly, binary masks were created for each granule to indicate its valid, non-cloudy pixels. A Sequential Floating Forward Search algorithm then systematically identified the optimal minimum subset of images based on the binary masks. For 26% of the tiles, a single image was already sufficient to achieve full tile coverage without cloud contamination. However, an average of 3.16 granules was required for the remaining tiles, with at least five granules needed for 38%. The compositing process itself was realised by pixel-wise retrieving the values from the granules in the subset that had the lowest blue band value at that specific pixel location.

In their publication accompanying the provided dataset, Kempeneers and Soille, 2017 explore the use of the granules' RGB quick looks. This approach requires the processing of only a spatially and spectrally reduced subset of granules. However, the possibility of scaling the processing to all spectral bands and their full spatial resolution is outlined. Even though

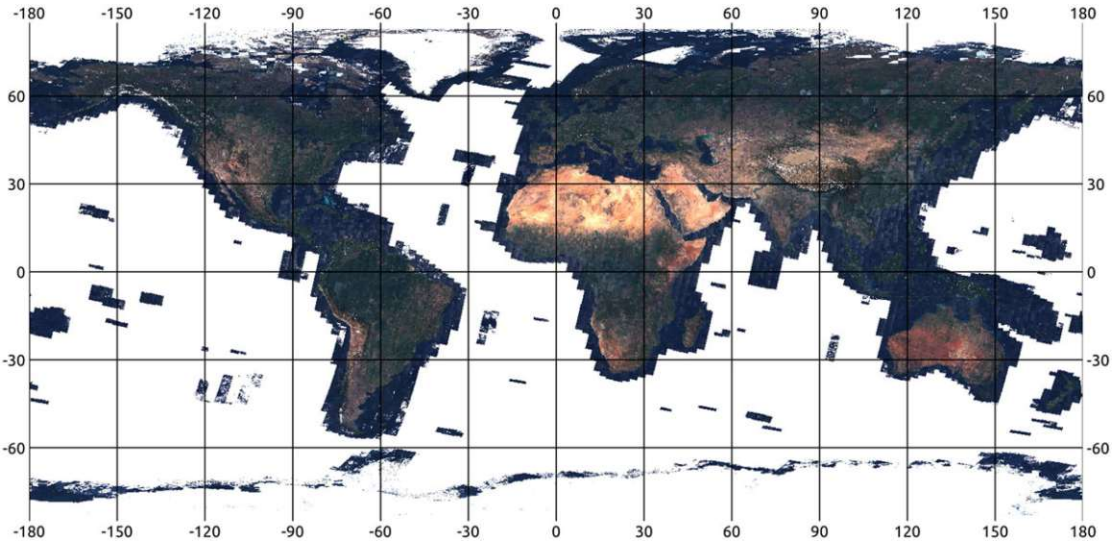

Figure 3.3: Global cloud-free true color composite in latitude longitude domain (taken from Kempeneers and Soille, 2017)

not mentioned explicitly, this evidently was done, as the provided datasets, available for download from the European Commission Joint Research Centre Data Catalogue (European Commission, Joint Research Centre, online), include composites of all spectral bands at the regular resolutions of the L2A granules.

For each of the 29,782 processed tiles, not only are the image files available, but also index rasters and file lists are available to indicate which pixel value was taken from which input file. Figure 3.3 shows the merged global RGB composite.

It has to be noted that because of the compositing algorithm, these datasets should be used with caution. Since pixel values of the output images are independently copied and not aggregated by mean or mode, over multiple granules, values within an image may stem from single observations made over a year apart or under different scene conditions. This compositing method often yields undesirable pixel patterns, possibly disturbing further analysis. However, regarding data volume estimation, exact values are not of prime importance, in contrast to completeness.

As datasets for conducting the experiments, the visible red SR band and the SCL (hereinafter referred to as reflectance and classification) were selected at both 10m and 20m samplings. The available 60m datasets could not be used as their sampling is incompatible with the 100km tile size of the E7Gs T1 tiling, which is a requirement for the estimation of the data volume reference values. Since reflectance bands diverge from classified bands in their scale of measurement, they are represented by different data types and storage capacity requirements. Additionally, the varying spatial sampling results in a different number of pixels per tile, affecting the file size as well. Overall, this choice of datasets yields four images per tile,

each featuring distinct characteristics and uncompressed file sizes, as can be seen in Table 3.1. These variations are expected to give insight into how the data itself and applied file compression affect the results. To provide a first overview of the data volumes involved, the actual data volume of the 10m reflectance dataset across all tiles is listed additionally.

Value type	Sampling	Bit-depth	Single tile...		Global volume
			Pixels	Raw vol.	
	[m]	[byte]	mio.	[MiB]	[GiB]
Reflectance	10	2	120.6	230.0	4307.5
Reflectance	20	2	30.1	115.0	1084.7
Classification	10	1	120.6	57.5	73.9
Classification	20	1	30.1	28.7	26.9

Table 3.1: Properties of the chosen datasets. Global volumes relate to the total of all downloaded 10m reflectance S2M tiles

Figure 3.4 exemplarily displays the 20m images for the S-2 tile 33UWP covering parts of Lower Austria and the city of Vienna in the east. Marked by the purple arrows, a clear edge is visible in both images. Investigating the corresponding index file reveals that, indeed, values on the western side stem from the beginning of April 2017, while pixel values east of the edge were acquired at the end of May 2017 from a different relative orbit. Additionally, inset maps offer a more detailed look at both image types at a close-to-pixel level. Differences to the corresponding 10m images only become visible when interactively overlaying them to the here displayed 20m images, thus their depiction is omitted.

3.3.2 Natural Earth land surface

In addition to the S2M datasets, a scheme to distinguish between land and non-land areas is essential, as only land areas are the object of investigation. For this purpose, the 1:10 million scale land polygon dataset from Natural Earth (Kelso and Patterson, 2010) was selected. This vector dataset, projected in the geographic latitude/longitude domain, provides a comprehensive global land surface definition and is freely accessible in shapefile format (Natural Earth, online).

Given its global coverage and simplicity, this dataset also served as the base dataset in the spatial overhead experiments. Directly aligning with this, the use of the Natural Earth land definition for the data volume overhead experiments ensures consistency in land delineation throughout the analysis and comparison of results.

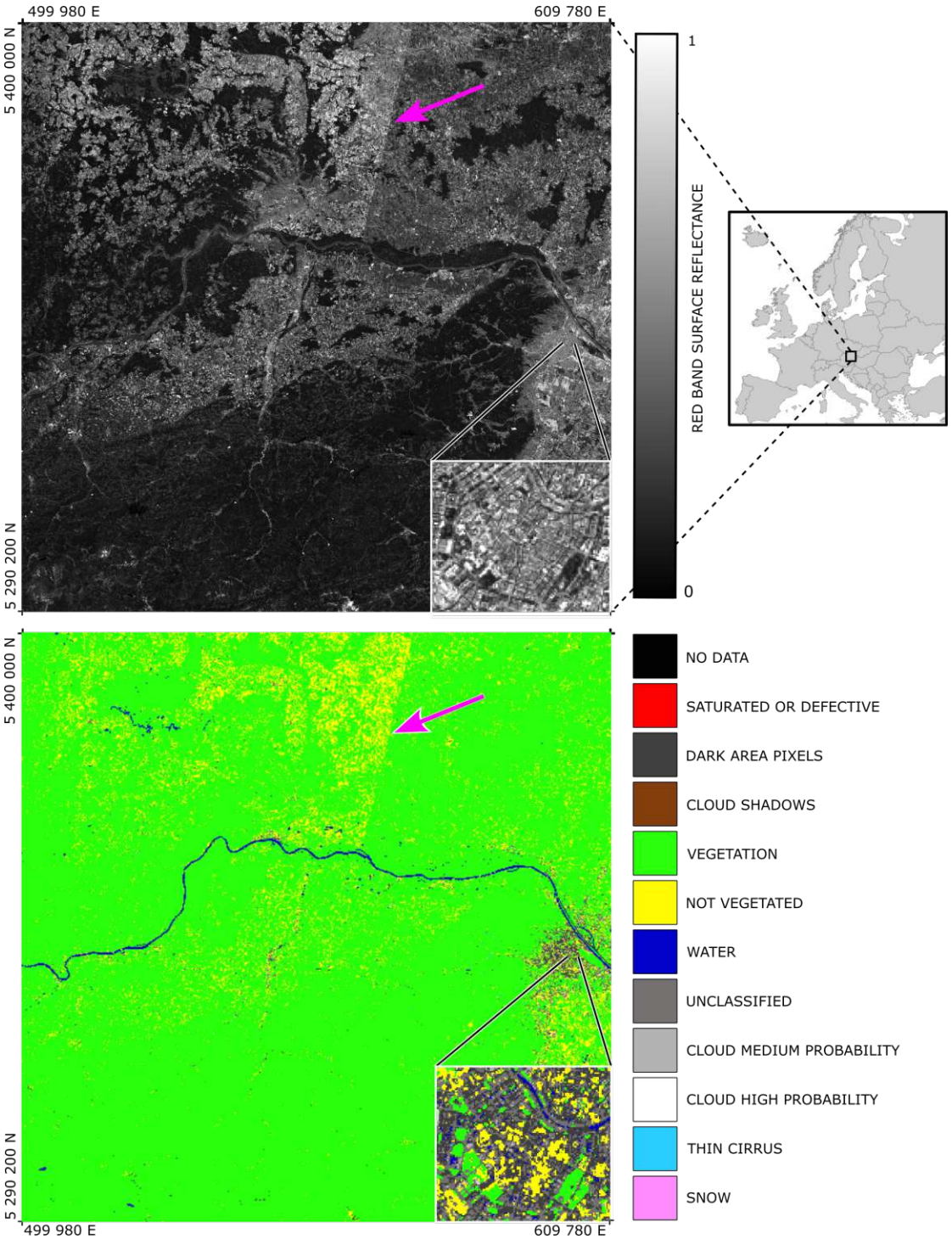

Figure 3.4: 20m reflectance (upper) and classification (lower) images for S-2 tile 33UWP (EPSG:32633, Unit [m]).

3.4 Data preparation

3.4.1 Land/non-land masking

While the S2M datasets fulfil all the necessary criteria for carrying out the volume overhead experiments, they cannot yet be used directly since they contain even more data than relevant for the scope of this study, which only includes land surfaces in the extent from 60°S to 85°N. Since the S-2 mission covers not only continental land areas and major islands but also the surrounding coastal waters, these are mapped in the original L2A granules and subsequently also exist in the S2M. Additionally, small parts of Antarctica are contained within the dataset, as can be seen in 3.3.

With regard to land surfaces, three types of tiles can be differentiated. Firstly, tiles that are entirely located within land surfaces. These only represent land pixels, making their respective images usable without processing. Secondly, tiles located entirely outside of land surfaces, which only contain non-land pixels. These are not of interest for the here conducted experiments at all and need to be sorted out. Thirdly, tiles located along coastlines, which contain both land and non-land areas, need to be masked in order to represent only land surfaces.

As a first step, tiles located entirely outside the set latitudinal limits are sorted out, as this reduces the amount of data to be processed. However, tiles intersecting the scope of interest at least partly were kept to ensure full coverage of the land surface is given. The previously described type of each remaining tile can then be identified through a spatial query using the tile's extent and the land surface dataset. However, every such query requires all data to be referenced in the same CRS. Thus, for each UTM zone, the land surface dataset was cropped to the zone's extent with a buffer to cover tiles extending beyond each zone's 6° width and reprojected to the zone's CRS. Application of these masks to the image data assigns a non-valid value to pixels outside of land surfaces. In case a masking result entirely consists of non-valid pixels, it carries no information on land surfaces and is sorted out. The quantity of the remaining images forms the S-2 version of the datasets, whose amount of redundancy is to be analysed.

Figure 3.5 depicts a global mosaic of the 20m reflectances merged from a) all downloaded S2M tiles and b) from all tiles of the S-2 version after the preparation steps, both in the lat/lon domain. In addition, tile 33UUB is used to demonstrate the masking process at a local scale in the native CRS of the tile (c).

It is important to note that the coastline observed in the imagery does not perfectly align with the land mask. This discrepancy means that some pixels visually resembling coastal areas remain included, while some land pixels are excluded at the same time. This is due to Natural Earth's coarser resolution compared to S-2, designed for generic cartographic use.

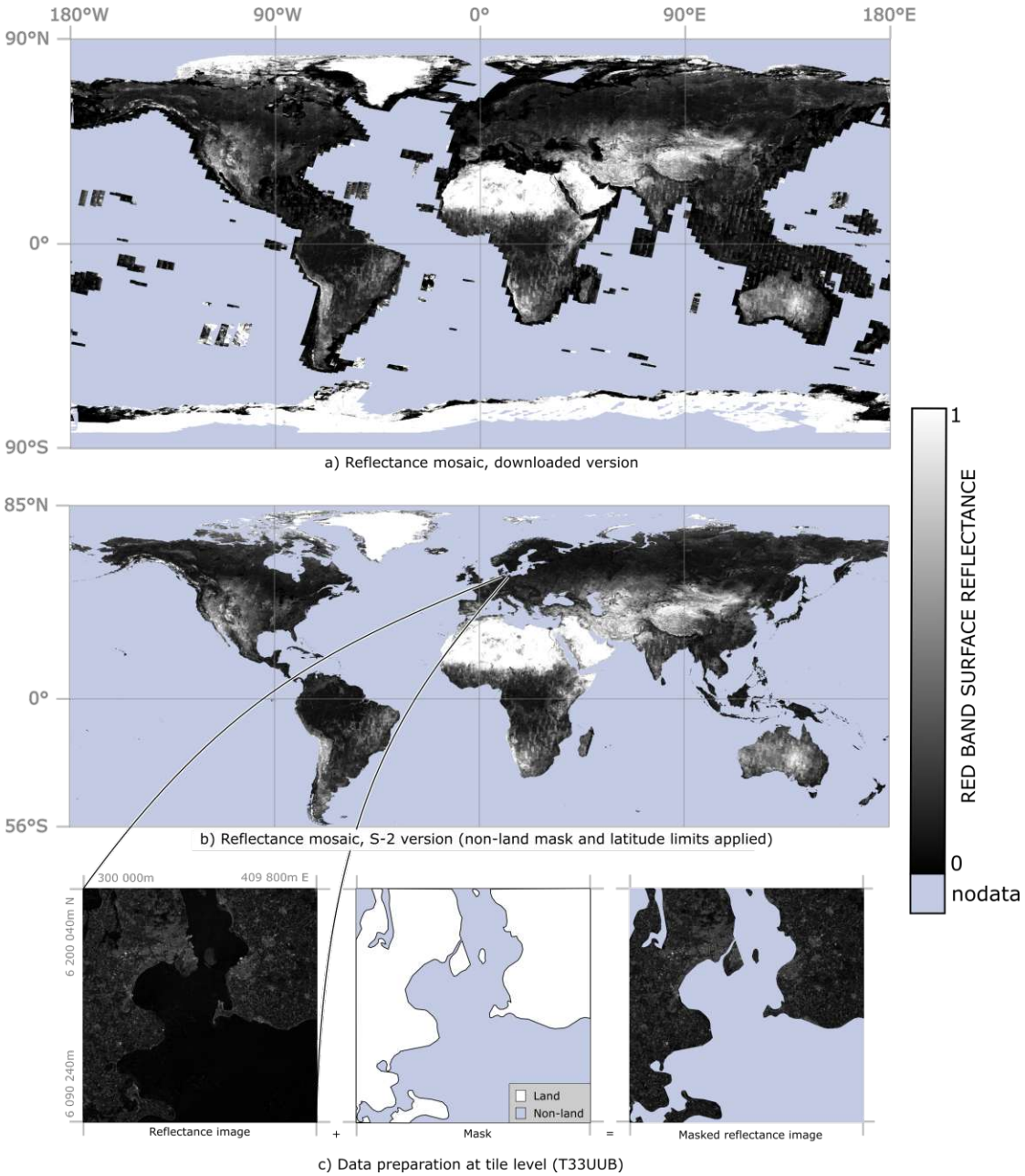

Figure 3.5: Reflectance mosaic from a) downloaded and b) prepared data, and c) masking at tile level (Tile 33UUB)

Overall, it is expected that the erroneous masking balances out over the entire examined domain. And, since this only applies to tiles along coastlines and exact pixel values are less critical to the data volume experiments compared to a consistent and complete land delineation, this discrepancy is accepted.

3.4.2 File format and conversion settings

So far, the data preparations were entirely conducted in the S-2 TGS, as this is the dataset's native framework. However, for the estimation of the reference values, the data is also required to be on hand in the E7G. For that purpose, each of the already masked and to the area of interest narrowed down data sets is converted to the E7G with the workflow proposed in Chapter 3.5. However, to ensure a meaningful comparison of data volumes, care must be taken to ensure that certain file format aspects are set reasonably and also appropriate methods during the conversion are applied to the different types of datasets.

The original S2M is stored as single-band GeoTIFF files. This format extends the standard Tagged Image File Format (TIFF) for storing image data with additional metadata for geo-referencing. Usually this includes information on the map projection, like an EPSG code, as well as an affine transformation that can be used to determine the projected coordinates of each pixel in the map plane (Bruce and Ruth, 2019). Thus, not all coordinates need to be stored explicitly, making this format ideal for the purpose of data volume analysis, and it is retained accordingly. The actual image data, i.e. pixels, are divided into groups forming either strips or rectangular blocks, which are compressed and stored separately. This means that only the corresponding groups and not the entire image array need to be fetched and decompressed to access individual pixels or image parts.

All data is stored in blocks of 256x256 pixels with Lempel-Ziv-Welch compression applied, both representing default GeoTIFF settings. Even though higher compression rates could be achieved using more complex compression algorithms or additional compression predictors, these also require higher processing power and more time for encoding and decoding data. The reflectance images are stored with two bytes per pixel since they contain reflectances in the form of integers with a validity range from 0 to 10000. For the classified images, one byte per pixel suffices, as these only feature 12 discrete classes. However, both image types use a value of zero to indicate masked, non-valid pixels. Apart from the georeferencing information, the image files do not contain any overviews or other metadata. The listed specifications are applied to all files of all datasets in all versions, except for the already explained sparse option, which is employed only for the reference version.

For both the TGS conversion procedure and the zonal mosaic creation, the ambiguity of multiple pixel values representing the same geographical locations due to overlapping tiles must be solved. To maintain the data quality at the highest level possible, for the reflectance

images, the arithmetic mean is calculated from all valid corresponding input values, while casting the result to integers. Regarding the classified data, the result is obtained by minimum aggregation in order to preserve the existing classification metric, likewise excluding masked pixels. Similarly, suitable resampling techniques must be selected in order to uphold data quality during the reprojection of the data from the UTM to the E7G spatial domains. Bicubic interpolation was employed for reflectances, which fetches a 4x4 source pixel neighbourhood for the calculation of a target pixel value. Meanwhile, to preserve the existing classes of the classified images, the nearest pixel value in the source domain was chosen.

3.5 Data conversion between Tiling Grid Systems

This section discusses the handling of image data with TGSs, covering the ingestion and extraction of single images, as well as the conversion of entire datasets between two TGSs.

3.5.1 Data ingestion

Ingesting a single georeferenced image, such as a preprocessed scene, into a TG is a common task when working with such frameworks. In a first step, all tiles of the TG that intersect the given source image are identified by transforming the source image's extent to the underlying CRS of the TG. By specifying the desired sampling of the output imagery, the pixel coordinates of each previously identified target tile are defined unambiguously due to the TG definition. Each target tile is then generated in an independent warping process. This image warping involves the reverse transformation of the target pixel coordinates into the CRS of the source image in order to retrieve the corresponding source values. These are then propagated forward to the target image. However, since the transformed pixel coordinates do not usually correspond precisely to the pixel coordinates of the source image, the forward-propagated values are determined by a resampling method. In the simplest case, the spatially closest value is chosen, or in more complex cases, a new value is interpolated from multiple neighbouring pixels. Target tile pixels located outside of the source image extent are set to be invalid, since no corresponding source values exist.

Ingesting an image into a TGS works similarly, with the difference that the source image possibly intersects tiles belonging to different TGs. Therefore, the extent of the source image must be transformed multiple times to identify all target tiles across all subordinated TGs. However, the warping procedure remains the same, as it is executed independently for each target tile in any case.

3.5.2 Data retrieval

Retrieving an image in an arbitrary projection from data stored in a TG is the inverse procedure to the just-explained image ingestion. First, the desired target extent is transformed to the underlying CRS of the TG to identify all tiles potentially containing relevant data. Subsequently, the data of these source tiles is merged into an image mosaic that covers the entire target extent. While merging tiles of TGs with non-overlapping tiles requires only the correct concatenation of the image arrays, overlapping tiles potentially demand the handling of multiple existing values per mosaic pixel. However, since referring to the same underlying PG, the arising ambiguities can be solved simply and efficiently by pixel-wise aggregation. The generated image mosaic is then warped to the pixel coordinates of the target image, which are defined by the desired sampling and extent. In certain warping scenarios, namely when target pixels coincide with mosaic edges, the determination of the forwarded pixel value may be affected by missing or unevenly distributed input values. To prevent such cases, the target extent is buffered before the identification of the source tiles. Likewise, the image mosaic is generated to span a larger extent in order to achieve sufficient coverage. The issue of appropriate input values for resampling is also the reason necessitating the intermediate mosaicking step, as the identical problem occurs along source tile borders when warping them individually, leading to quality-degrading artefacts in the target image.

If an image is to be retrieved from a TGS, the target extent might be partly within the coverage of multiple TGs. Accordingly, data from all these source TGs is required to build the target image. However, not all source tiles may be combined into a single mosaic since their underlying PGs differ. Necessarily, the retrieval procedure is applied to each contributing TG individually. Each resulting target image might then be only partly filled with valid data if the corresponding TG does not fully cover the target extent. However, due to the standardised procedure and the preset target pixel coordinates, all resulting target images are of equal extent and aligned with each other. This enables the stacking into a data cube, with each layer representing the retrieved data of a single source TG. Target pixels that are assigned only one valid value across all layers are located in areas only covered by a single source TG. In contrast, pixels with several valid values in different layers represent areas covered by multiple TGs. Applying a pixel-wise aggregation function to the data cube solves potential ambiguities while implicitly forwarding the exact values in areas covered by a single TG, ultimately resulting in the sought single-layer output image.

3.5.3 Data conversion

Converting entire datasets between TGSs is essentially nothing other than repeatedly executing the image retrieval process. In doing so, the tiles of the target TGS specify the desired image extents and CRSs. However, a large number of target tiles may accumulate, depending

on the geographical coverage of the source dataset. Therefore, not only a quality-preserving but also runtime-efficient implementation of the conversion is essential. The following workflow was developed and implemented to convert data between the S-2 TGS and the E7G in both directions. Accordingly, the accompanying conversion example (fig. 3.6 also aims to generate an E7G tile from the S-2 TGS. However, the approach is applicable to convert data between all TGSs as defined in the introduction.

The prerequisites of the conversion implementation are the definitions of both involved TGSs. These include the definitions of the map projections and coverages of all subordinated TGs. Furthermore, for each TG, the geometries of the associated tiles must be detailed. Another requirement is the compatibility of the data sampling and the target TGS, meaning that the target tile size must be an exact multiple of the data sampling. Otherwise, pixels would protrude beyond the tile border or not completely fill it. When leaving the scope of a single tile, this yields non-aligned images, as adjacent tiles then do not refer to the same underlying PG anymore, contradicting the basic concept of TGs. Hence, compatibility must be checked beforehand. In principle, altering the data sampling during conversion is possible as long as the above condition is met. However, the purpose of this workflow is to transfer the data to another system while retaining its quality, enabling equal usability in any TGS. Accordingly, up- or downscaling is not dealt with here.

Apart from fulfilling the prerequisites, the starting point of the TGS conversion workflow is the listing of all desired target tiles. Subsequently, the image retrieval process outlined previously is carried out for each target tile individually. The specific processing steps are listed in the following. Additionally, Figure 3.6 provides a visual example of how the E7G Tile EU_E054N015T1 is retrieved from S2M 20m red band SR data stored in the S-2 TGS.

1. In order to identify all intersecting tiles of the source TGS, a spatial query is performed based on the target tile geometry. To speed up this query, the subordinated source TG coverages are checked for overlaps first. This filtering reduces the number of checks required at tile-level, as non-intersecting source TGs cannot contain intersecting tiles. However, any kind of spatial query requires all involved geometries to be referenced to the same CRS. Since the underlying projections of TGSs only cover specific parts of the Earth's surface, neither source nor target CRSs are suitable to be used as query domain. Per se, reprojecting a geometry from one projected domain to another, covering an entirely different part of the Earth's surface, works coordinate-wise. However, it easily leads to topological misinterpretations and, subsequently, erroneous query results. Accordingly, all geometries are transformed to the geographic latitude/longitude domain. While not suitable for raster processing at TG level due to its heavy distortions, this projection with global coverage is suitable for conducting vector-based spatial queries. Only errors due to the antimeridian discontinuity may arise when geometries are mapped onto it. In this case, the geometry's shape is split into two parts, which are

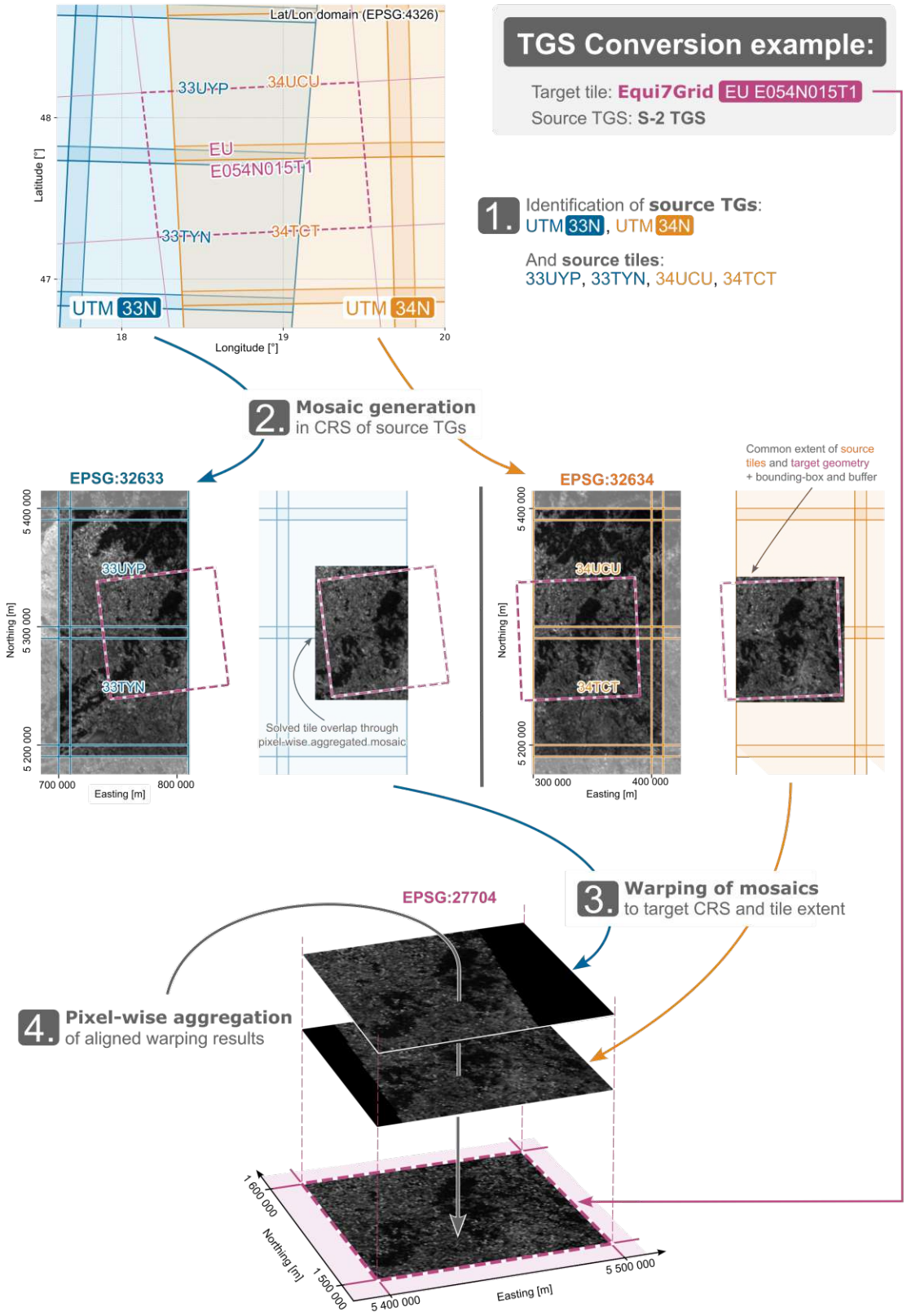

Figure 3.6: Exemplarily retrieving E7G tile EU_E045N015T1 from data stored in the S-2 TGS

handled independently. Accordingly, conducting the queries in the latitude/longitude domain reliably returns all sought tiles of the source TGS already sorted into groups of equal TGs.

2. For each group, an individual image mosaic is created from the corresponding image files, as described in the image retrieval process. However, the joint coverage of a tile group usually exceeds the extent of the target tile, resulting in the processing of non-required data. To prevent unnecessary processing, the target tile geometry is transformed to the CRS of the tile group in order to determine the exact needed mosaic boundaries. Due to the change in the spatial domain, the geometry is usually no longer represented by an axis-parallel rectangle as in its native projection, but a commonly rotated four-sided polygon. Using this polygon as the boundary for the mosaic thus requires the application of non-performant geometric cutline operations to determine which pixels of the source tiles are inside and which are outside this intersection. Therefore, the polygon is extended to axis-parallelity using a bounding-box operation. This simplification admittedly carries along redundant data, but enables the mosaic generation with array indexing for efficient reading and writing operations to benefit runtime. Lastly, to prevent the earlier-mentioned resampling issue of missing or not evenly distributed input data during the later performed warping, the bounding box is additionally buffered to all four sides, providing sufficient data.

   Once the final mosaic extent is determined, the respective pixel subsets are extracted from the source tiles and concatenated. To mosaic pixels that are assigned multiple values from different tiles, pixel-wise aggregation functions are applied to determine a single representative value. For example, reflectances or backscatter may be averaged while discrete classifications or binary masks require the selection of the minimum, maximum or mode value to preserve their meaning. In general, the mosaicing process results in a single mosaic for each involved source TG within its native CRS. These mosaics cover the target tile geometry with as little redundant data as possible without requiring runtime-costly algorithms.

3. In the next step, all mosaics are warped to the CRS and pixel coordinates of the target tile, though still treated separately. Thus, the warping outputs are jointly georeferenced as well as aligned with each other. This facilitates easy merging by stacking the warped images on top of each other, forming layers of a data cube. The data cube represents the target tile, with each layer referring to a source TG. Depending on whether a source TG covers the target tile partly or wholly, the corresponding layer might only partially hold valid data. However, due to the gaplessness of a global TGS, all pixels of the datacube are covered at least once across all layers, while pixels represented by several source TGs are assigned several values.

4. Accordingly, a pixel-wise aggregation function is applied to combine all layers into a single image, which is the final result of the conversion. While the mosaicking step earlier solved potentially overlapping tiles within the source TGs, the datacube reduction performed here solves potential overlaps of tiles belonging to different TGs. Thus, the same type of aggregation function is used, depending on the data. In the typical case of a target tile being entirely located within a single source TG, only one mosaic is created, and its warping output is already the final conversion output.

Overall, this workflow takes all available data from all available source images into account to maintain data quality, while particularly optimising the query and mosaicking processes for run-time efficiency. The pure warping procedure and datacube operations are taken over by dedicated software packages, thereby already are most efficient.

3.6 Tools and software

All data handling and processing tasks done in this thesis were performed using the high-level Python programming language (version 3.10.6). As an open-source environment, Python provides access to a vast environment of powerful libraries tailored to a broad range of applications, including geospatial data management and image processing.

Foremost among these is the Geospatial Data Abstraction Library (GDAL, version 3.5.2), which serves as a foundational toolkit for handling raster and vector data formats, together with the integrated OGR Simple Features Library. Developed in C/C++ for high-performance computing, GDAL also offers Python bindings, enabling seamless integration within Python scripts. In this thesis, GDAL serves as the principal tool for reading and writing geospatial image data. Particular emphasis is placed on the gdalWarp utility, which facilitates the transformation of raster datasets between different CRSs through reprojection and resampling. Additionally, it allows for the use of vector data to mask and extract pixel values or crop rasters based on spatial constraints. The tool also supports the creation of image mosaics, enabling the merging of individual tiles to build larger spatial composites (GDAL/OGR contributors, 2025).

While GDAL handles the geospatial processing layer, the NumPy library (version 1.23.4) is responsible for managing numerical array operations. NumPy provides fast and efficient manipulation of large data arrays through advanced indexing, vectorisation, and a comprehensive suite of mathematical functions. Also written in optimised C code, NumPy combines Python's flexibility with the performance of compiled languages. In the context of this work, the library was used for tasks such as aggregating image stacks through pixel-wise calculations or concatenating spatially adjacent raster arrays (Harris et al., 2020).

To manage complex vector datasets, such as global tiling schemes composed of numerous individual geometries, the GeoPandas library was employed. GeoPandas extends the functionality of the widely used Pandas library to support spatial data types. This enables operations such as spatial joins, intersection tests, and buffering, which are crucial for spatial queries and geometric data manipulation throughout the study (Jordahl et al., 2020, July).

4 Results and discussion

After downloading the S2M datasets, the necessary preparation steps were carried out. This involved filtering out tiles located beyond the defined latitudinal boundaries as well as applying a land mask to exclude non-land areas. The outcome of this process are the S-2 dataset versions, which form the basis of the subsequent analysis. Conversion of these datasets into the E7G framework produces the E7 versions. By further restricting these converted datasets to the unbuffered extent of the E7G, the reference versions are obtained, from which the reference data volumes can be derived. Although the E7 versions primarily serve as an intermediate step for determining the reference values, they also allow for quantifying the data volume overhead introduced by the E7G itself. Accordingly, their data volumes are likewise determined and included in the analyses. An overview of all dataset versions used in this study is provided in Table 4.1.

				Nr. of tiles:	
Version	TGS	60°S-85°N	Only land-covering	Expected	Existing
Downloaded	S-2	False	False	/	28,215
S-2	S-2	True	True	18,624	18,357
E7	E7G	True	True	17,002	16,785
Reference	unbuffered E7G	True	True	16,897	16,709

Table 4.1: Dataset versions and their properties

The term *expected number of tiles* refers to the number of tiles that can be derived from the respective TGS definitions through spatial queries, when considering both land cover and the tile geometries. If a version includes exactly one image per dataset per expected tile, it can be considered to achieve full coverage of the major global landmasses, excluding Antarctica. This assumption can be verified by contrasting the expected count to the actual number of files present in each dataset version, listed as *existing number of tiles*. However, this comparison reveals a small shortage in tile counts for all dataset versions. This discrepancy is attributed to smaller islands that are included in the Natural Earth land definition but are located outside the Sentinel-2 mission's coverage. A visual inspection confirms this explanation and verifies that the discrepancy is not due to missing data. Thus, all dataset versions successfully achieve the intended land surface coverage, despite the numerical gap.

A comparison of the tile counts listed for the S-2 and E7 versions relative to the reference version already provides an initial indication of the system's redundancies, even without considering data volumes. Although the tile size of the S-2 TGS is nominally larger, their overlaps result in effectively the same tile sizes of 100km for both systems, enabling a direct comparison. Within the spatial domain spanning 60°S to 85°N, the S-2 version requires nearly 10% more tiles than the reference version to achieve complete landmass coverage. In contrast, the E7G introduces less than 1% additional tiles. This difference in efficiency can

be attributed to the structural designs of the systems. The E7G strategically positions zone boundaries in a way that largely avoids intersecting landmasses. As a result, overlaps of tiles occur primarily over oceans, making the associated redundancies inconsequential for land surface datasets. Therefore, land areas are almost exclusively covered by single tiles within a single TG. This efficiency is further supported by the fact that the tile placement within a TG follows the regular structure of the corresponding tile matrix. Thus, larger TGs naturally result in more favourable tile arrangements, as fewer boundaries are introduced in general. Comparably, the S-2 TGS relies on a larger number of smaller zones. These are defined purely geometrically and do not account for the global distribution of landmasses. As a result, tile overlaps occur not only more frequently in general, but also more frequently over land areas. Added to this is the greater variety of tile matrices with non-aligned structures, which further contributes to overall less favourable and more redundant coverage.

4.1 Global, non-polar domain

The first experimental setup considers the total data volume of the datasets within the latitudinal bounds of 60°S to 85°N. As this setup encompasses the largest spatial extent among all configurations analysed in this study, the data preparation process was explicitly designed to produce this exact coverage. Consequently, no further selection or filtering is required for any of the dataset versions. Summing the file volumes of each version directly yields the absolute values presented in Table 4.2.

	Data volumes [GiB]+ Volume overheads [%]		
Data set	Reference	Equi7Grid	S-2 TGS
10m reflectance	2133.0 (=100)	2158.0 +1.2	2809.7 +31.7
10m classified	38.3 (=100)	38.7 +1.0	47.1 +22.9
20m reflectance	539.4 (=100)	545.8 +1.2	711.1 +31.8
20m classified	13.5 (=100)	13.7 +1.0	17.4 +28.9

Table 4.2: Measured data volumes in the global, non-polar domain

Due to the differing characteristics of the datasets, the resulting data volumes vary significantly in size, ranging from several thousand to only a few GiBs. Therefore, Table 4.2 also lists the relative data volume overheads, while Figure 4.1 provides a graphical representation. This enables a standardised comparison across the variety of datasets, as well as with the additionally indicated findings from the spatial overhead experiments.

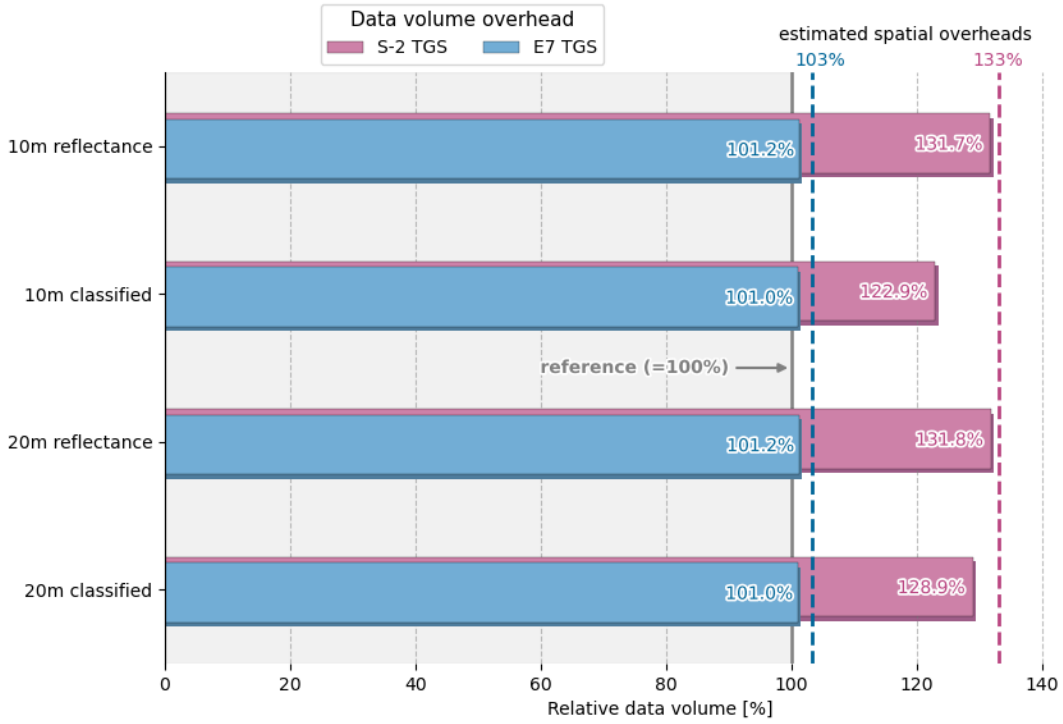

Figure 4.1: Data volume overheads measured for the global, non-polar domain

For the 10m reflectance dataset stored in the S-2 TGS, the observed volume inflation of 31.7% lies slightly below the previously derived spatial overhead of 33% for the non-polar domain. In the case of the E7G, a spatial overhead of 3.2% was reported. However, this value refers to the global domain. Assuming that it slightly decreases when applying latitudinal constraints, thus reducing the mapped area, the observed volume inflation of 1.2% of the E7 version remains consistent with spatial overhead in magnitude. A comparison with the results from the 20m reflectance dataset reveals that the relative volume inflation remains exactly the same for the E7G and increases only marginally to 31.8% for the S-2 version. Since the only difference between the 10 and 20m reflectance datasets lies in their spatial resolution, the slight variation in inflation can be attributed to the differing pixel counts. Given that a higher number of pixels results in a lower volume overhead, it can be concluded that this effect is driven by image compression. With a greater amount of pixels, more data is passed to the compression algorithm, leading to a higher compression ratio and therefore less relative volume increase. However, since the observed difference is non-significant, the spatial resolution apparently has only minimal impact on the data volume overhead.

For the classified datasets, an entirely different scenario emerges. The volume overheads for the S-2 TGS contrast significantly across the spatial resolutions, amounting to approximately 23% and 29%, respectively. These figures deviate not only from each other but also from the corresponding results of the reflectance datasets. In contrast, the E7G volume inflations

remain consistent at 1.0% for both resolutions and are also in line with the values observed for the reflectance data. These results are primarily attributable to two contributing factors:

- The first factor lies in the fundamental difference between reflectance and classification data, which inherently require different amounts of storage per pixel, as outlined in Table 3.1. The reflectance datasets are stored using two bytes per pixel, allowing for an integer value range from 0 to 65,535. However, since reflectance values are only valid within the range of 0 to 1 by definition, they are scaled during storage using a quantification coefficient, mapping them to a range from 0 to 10,000. As a result, only around 15% of the possible values defined by the data type are actually utilised. In contrast, the classified datasets are stored using a single byte per pixel, corresponding to a possible integer value range from 0 to 255. The S-2 SCL definition, however, defines only 12 valid classes, meaning that merely about 5% of the available values are in use. Thus, the pixel values in the classified datasets are naturally more homogeneous than those in the reflectance datasets. In this context, it should be taken into account that achievable compression ratios depend on the variance of the jointly stored pixel values, as similar values can be encoded more efficiently. Accordingly, the classified datasets exhibit lower data volume overheads when compared to the reflectance ones.

- The second factor, also rooted in the characteristics of the data itself, accounts for the significant differences in volume overhead between the S-2 versions of the 10m and 20m classified datasets. The SCL is not natively produced at a 10m resolution, as the entire L2A classification process is carried out at a 20m sampling (ESA, 2021). Accordingly, when downloading an L2A granule, the SCL is only available at a 20m sampling. However, the S2M data used here includes a 10m version of the SCL, although no details about its generation are provided in the documentation. Therefore, it must be assumed that the 10m dataset was derived by resampling the native 20m image of each tile. This would imply a simple upscaling by a factor of two, where each native 20m pixel exactly matches a 2×2 block of 10m pixels. Given the classified data type, a class-preserving resampling technique must have been applied, meaning that every 2×2 pixel neighbourhood uniformly inherits the classification of its parent 20m pixel. As a result, while the 10m images contain exactly the same informational content as their 20m counterparts, it is distributed across four times as many pixels. These artificial duplicates lead to exceptionally high compression rates during storage, as the repeated class values form homogeneous pixel groups that can be efficiently encoded by the compression algorithm. The duplicate effect does not apply to the reflectance datasets, which are resampled from their native acquisition geometries to the respective spatial resolutions during Level-1 processing and then treated independently. While the neighbourhoods of identical 2×2 pixel blocks are not re-

tained during the conversion from the S-2 TGS to the E7G, the data still originates from a native 20m sampling. Thus, it does not genuinely represent spatial structures at a 10m scale. Accordingly, the 10m classification results should be interpreted with caution and excluded from direct comparisons when assessing the data volume overheads across different TGS. Nevertheless, for the sake of completeness, they are still included in the following experiment setup. Unfortunately, this does not allow for verifying the assumption that different spatial sampling does not influence the data volume overhead of otherwise identical data sets.

Compression analysis

Overall, the previously described results of this first experiment setup suggest that the observed data volume overheads are strongly influenced by the achievable compression rates of the respective datasets. In order to assess whether different TGS and thus different dataset versions are compressed at varying extents, potentially distorting the comparison, a simple analysis was conducted. This combines the information from Tables 3.1 and 4.1 to estimate the achieved compression ratios per dataset and version. By multiplying the raw, uncompressed file size of a single image by the number of tiles in the corresponding data version, an estimate for the global uncompressed data volumes can be obtained. Comparing these theoretical volumes to the actual, measured data volumes listed in Table 4.2 yields the compression ratios reported in Table 4.3.

The previously observed patterns across the different datasets are consistently reflected in the compression ratios for all versions. First, the 10m reflectance datasets consistently achieve minimally higher compression ratios than their 20m counterparts. This supports the earlier hypothesis that, when comparing datasets that differ only in spatial resolution, passing a larger amount of data to the compression algorithm results in relatively more efficient compression. Second, the 10m classification datasets show significantly better compression ratios than the 20m versions, which is an expected outcome of the 2×2 pixel duplicates introduced during upscaling. Third, the differences between reflectance and classification datasets clearly reflect the varying degrees of variance caused by the different data types and their value ranges.

When comparing the compression ratios across the different data versions, the values are found to be of consistent magnitude. This indicates that the compression ratio of a dataset is largely independent of the TGS in which it is stored, even though the dataset is spatially differently tiled. Consequently, the methodology employed in this thesis is further validated as it implicitly assumes equal compression behaviour across all versions of a dataset. Accordingly, the observed differences in data volume refer to the different redundancies the TGSs induce when mapping the Earth's land surfaces.

Version	10m refl.	20m refl.	10m class.	20m class.
S-2				
Tile raw volume [MiB]	241.1	60.3	120.6	30.1
Nr. of tiles		18,357		
Global raw volume [GiB]	4322.5	1080.6	2161.3	540.3
Global actual volume [GiB]	2809.7	711.1	47.1	17.4
Compression ratio	**1.54**	**1.53**	**44.95**	**30.90**
E7				
Tile raw volume [MiB]	200.0	50.0	100.0	25.0
Nr. of tiles		16,785		
Global raw volume [GiB]	3278.3	819.6	1639.2	409.8
Global actual volume [GiB]	2158.0	545.8	38.7	13.7
Compression ratio	**1.52**	**1.50**	**42.36**	**29.91**
Reference				
Tile raw volume [MiB]	200.0	50.0	100.0	25.0
Nr. of tiles		16,709		
Global raw volume [GiB]	3263.5	815.9	1631.7	407.9
Global actual volume [GiB]	2133.0	539.4	38.3	13.5
Compression ratio	**1.53**	**1.51**	**42.60**	**30.22**

Table 4.3: Estimated compression ratios

The slightly higher compression ratios observed for the S-2 versions compared to the E7 versions can be attributed to the larger total data volumes available to the compression algorithm. Since the E7G consists of fewer tiles and thus pixels overall, there is comparatively less data to compress. The fact that the reference versions show slightly higher compression ratios than the E7 versions, despite being nearly identical apart from the masking or removal of certain tiles, is explained by the use of the sparse GeoTIFF option for the reference datasets. This specification allows entirely empty pixel blocks to be skipped thoroughly during storage rather than being simply compressed. However, given that the improvement in compression ratios is only minimal, it can be concluded that the sparse option does not yield significant advantages. Even the here used standard compression method appears to handle completely empty pixel blocks with sufficient efficiency.

4.2 Continental domains

The second experimental setup examines system-induced redundancy at the continental scale, while still confined to the latitudinal range of 60°S to 85°N. Accordingly, the same analysis described in Section 4.1 was applied individually to each continent, excluding Antarctica,

which lies entirely outside the defined spatial domain. To estimate the data volumes, only those tiles that intersect with the respective continent were considered. This was achieved by performing spatial queries in both TGSs to identify the relevant tile subsets. Summing the corresponding file volumes yields the continental data volumes, which are presented in Table 4.4 for each dataset, version, and continent.

A preliminary observation reveals that the sum of all continental data volumes for a given dataset and version exceeds the global totals listed in Table 4.2. This discrepancy arises from the fact that tiles located along the defined continental boundaries are assigned to all relevant continents, resulting in multiple counts. Such multiple assignments are necessary, as complete coverage of an area requires to include all tiles that intersect it, even if only partially.

Another important observation relates to the E7 version volumes for Oceania and South America, which match their corresponding reference volumes exactly. This outcome is attributable to the specific continental zone definitions applied in this thesis, which do not align with internal boundaries of the E7G. For example, the overlapping region between E7G-Asia and E7G-Oceania falls entirely within the continent of Asia as defined here. As a result, Table 4.4 depicts overhead values of 1.7%, for Asia, while Oceania exhibits no tile overlaps at all and therefore matches the reference version exactly, displaying 0% overhead. Consequently, interpreting the E7 volume overheads in this experimental setup is not meaningful, as the results are directly dependent on the specific choice of zone boundaries. In order to maintain consistency with the spatial overhead experiments, the same continental definitions are used here. However, any modification to these would alter the results accordingly, but also compromise the comparison of the different metrics.

Accordingly, for this experiment setup, only the data volume overheads of the S-2 TGS are considered. These values are presented by continent in Figure 4.2, allowing the vastly differing absolute data volumes to be placed into a meaningful relative context. The same general pattern observed for the global, non-polar domain is consistently reflected across all continents. The 20m reflectance datasets feature only minimally higher inflations than their 10m counterparts. The 20m classification datasets show slightly lower inflation values compared to the reflectances, while the 10m classification values are found to be significantly lower.

Accordingly, the previously identified causes for these deviations apply not only to the global land surface but also to individual continental land surfaces. Therefore, the variations between continents become of higher interest in this analysis setup. As identified by Bauer-Marschallinger and Falkner (2023), the local redundancy of the S-2 TGS is dependent on the geographic latitude at which a continent is located. This finding is backed by the observed data volume overheads, which generally align with the pattern of the previously determined spatial overheads. For example, Africa and South America, both situated along the equator,

Dataset	Data volume [GiB] Reference	E7G	S-2 TGS	Overhead [%] E7G	S-2 TGS
			Africa		
10m reflectance	484.2	485.5	614.3	+0.2	+26.8
10m classified	6.8	6.8	8.3	+0.1	+21.7
20m reflectance	122.1	122.2	155.0	+0.1	+26.9
20m classified	2.4	2.4	3.0	+0.1	+25.1
			Asia		
10m reflectance	728.4	741.1	962.5	+1.7	+32.1
10m classified	14.5	14.7	17.5	+1.7	+21.0
20m reflectance	184.2	187.4	243.9	+1.7	+32.4
20m classified	5.1	5.2	6.5	+1.7	+28.4
			Europe		
10m reflectance	143.1	152.7	195.2	+6.8	+36.5
10m classified	2.6	2.7	3.2	+5.0	+24.1
20m reflectance	36.4	38.8	49.7	+6.7	+36.6
20m classified	0.9	1.0	1.2	+4.7	+30.6
			North America		
10m reflectance	383.9	384.5	527.8	+0.2	+37.5
10m classified	8.2	8.3	10.5	+0.1	+26.6
20m reflectance	97.3	97.4	134.0	+0.2	+37.7
20m classified	2.9	2.9	3.9	+0.1	+34.1
			Australia		
10m reflectance	138.5	138.5	179.6	+0.0	+29.7
10m classified	1.9	1.9	2.4	+0.0	+25.1
20m reflectance	34.8	34.8	45.3	+0.0	+30.0
20m classified	0.7	0.7	0.9	+0.0	+27.5
			South America		
10m reflectance	266.2	266.2	342.5	+0.0	+28.7
10m classified	4.4	4.4	5.4	+0.0	+22.1
20m reflectance	67.0	67.0	86.5	+0.0	+29.1
20m classified	1.6	1.6	2.0	+0.0	+25.2

Table 4.4: Measured data volumes and overheads of the continental domains

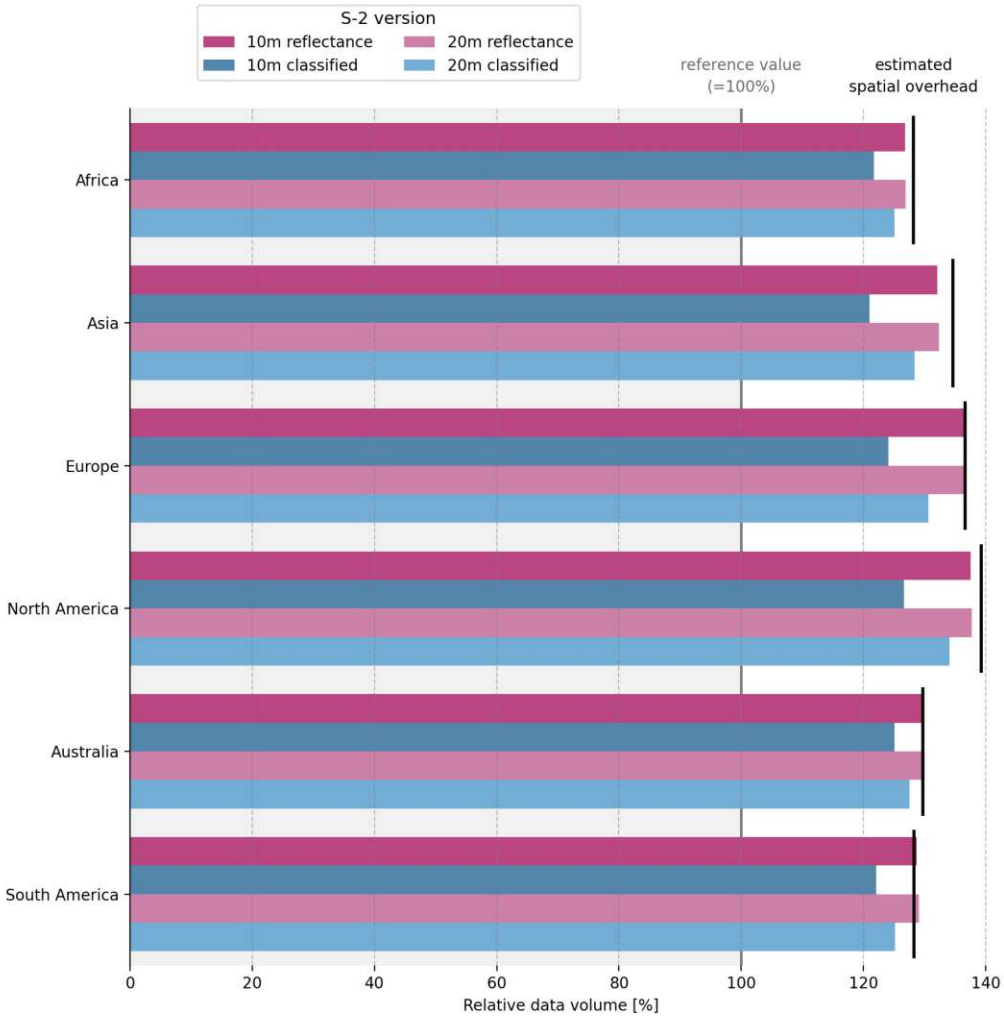

Figure 4.2: Data volume overheads of the S-2 TGS at continental level

exhibit comparably low redundancy values. In contrast, the highest overhead rates are found in continents located at higher latitudes, such as Europe and North America. The root cause lies in the structural design of the UTM system, which partitions the Earth's surface into longitudinal strips. At higher latitudes, this setup results in an increased density of overlaps between tiles from adjacent zones, thereby generating a higher volume of redundant data.

While these volume-based results generally reflect the expected latitudinal trend, some deviations emerge when examined per continent. For instance, Europe and Australia show a near-perfect match between the reflectance data inflation and spatial overheads. In contrast, South America exhibits slightly higher volume overheads, while Asia displays slightly lower ones. These discrepancies suggest that, beyond the latitudinal trend, another factor is affecting the data volume overheads. In the land surface experiments, only the raw number of

land pixels per continent was considered. In contrast, for real image datasets, compression efficiency is also affected by the variance of pixel values within each tile and tile-internal pixel block. This variance, in turn, is influenced by how and across how many tiles the underlying PG is spatially divided. Accordingly, Figure 4.3 presents the absolute number of tiles assigned to each continent in the frame of the analysis. Additionally, based on the known land areas of the continents (sourced from Bauer-Marschallinger and Falkner (2023)) and the fixed dimensions of S-2 tiles, a theoretical minimum number of tiles required to cover each continent can be calculated. This theoretical value, which assumes ideal and compact coverage while disregarding the actual geometric complexity of landmasses, is also indicated in the figure. The resulting ratio between actual and theoretical tile counts provides an indication of how efficiently tiles are distributed over each continent's landmass. For example, Africa, due to its compact and contiguous land shape, requires only about 5% more tiles than the theoretical minimum. In contrast, Europe and Australia require significantly more, 25 and 31% respectively, primarily due to their complex coastlines and the presence of numerous island groups. Such geographic features lead to the inclusion of many only partially land-covering tiles, splitting the data across a greater number of tiles. This fragmentation potentially increases variance within jointly compressed pixel blocks, ultimately resulting in higher storage requirements.

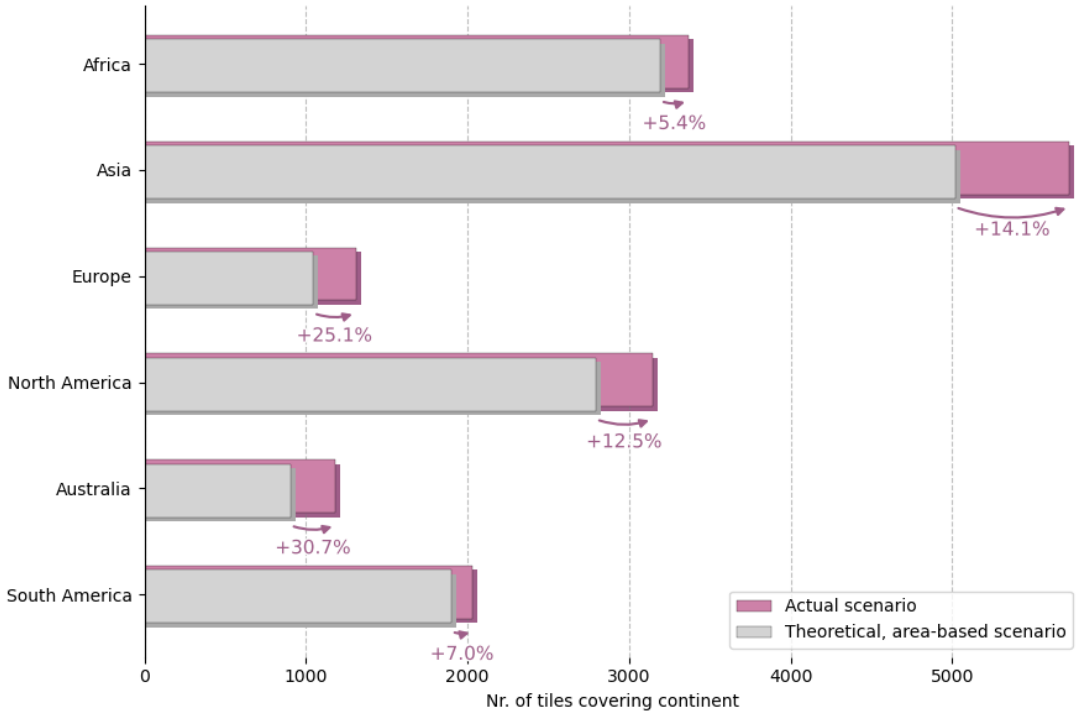

Figure 4.3: Tile analysis on continental level

Based on the hypothesis that the ratio between actual and theoretical tile counts is the primary factor driving the previously observed deviations, the volume overhead patterns are

expected to closely reflect these ratios. For example, in the case of South America, where the measured volume overhead exceeds the spatial overhead, this would imply a disproportionately high number of tiles. However, this is not the case, and more generally, no consistent correlation between tile count and data volume overhead emerged. This suggests that image compression effectively absorbs much of the structural inefficiency that might otherwise result from fragmented tile coverage. In essence, while the number and distribution of tiles influence how data is organised, the actual volume impact is strongly governed by the compressibility of the data itself.

4.3 Overlap type analyses

The third and final experimental setup aims to analyse the composition of redundancy within the S-2 TGS. Specifically, it aims to quantify how much of the additional data volume arises from overlapping tiles within individual UTM zones, and how much is attributable to overlaps between tiles from adjacent zones. To achieve this, all tiles assigned to the same UTM zone were mosaicked using pixel-wise averaging in areas of overlap in order to eliminate intra-zone overlaps. The data volumes of the resulting zonal files were then summed, and the total values are reported in Table 4.5. For reference and comparison, the volume values previously determined for the global, non-polar domain are used as benchmarks. The corresponding relative overhead rates are visualised in Figure 4.4.

	Data volumes [GiB]+ Volume overheads [%]		
Data set	Reference	Equi7Grid	S-2 zonal
10m reflectance	2133.0	2158.0	2376.0
	(=100)	+1.2	+11.4
10m classified	38.3	38.7	39.8
	(=100)	+1.0	+3.7
20m reflectance	539.4	545.8	601.2
	(=100)	+1.2	+11.5
20m classified	13.5	13.7	14.5
	(=100)	+1.0	+7.4

Table 4.5: Measured data volumes of the merged zones

It is important to note that mosaics were not created for all UTM zones, as some zones did not contain a single land-covering tile based on the criteria defined in this analysis. These zones were therefore excluded from the mosaicking process. Consequently, only 107 zonal mosaics per dataset were generated from the original 18,357 tiles, rather than the 120 that would have been expected if all zones had contained relevant land coverage.

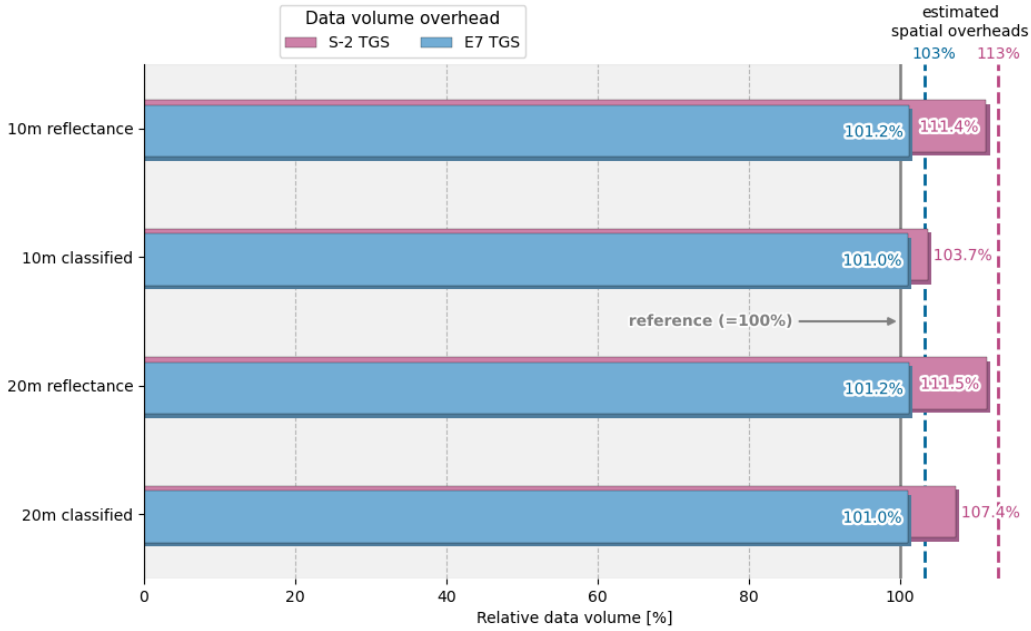

Figure 4.4: Data volume inflations of the S-2 TGS after merging zones

For the 10m reflectance dataset, the volume inflation is reduced to approximately 11.4% after mosaicking, compared to 31.7% in the original tiled S-2 version. Similarly, the 20m version decreases from 31.8% to 11.5%. These results indicate that approximately two-thirds of the total redundancy is attributable to intra-zone overlaps, with the remaining third arising from inter-zone tile intersections. A comparable pattern is observed for the classification datasets, as the 10m version shows a reduced inflation of 3.9%, down from its original 21.6%, and the 20m version drops from 27.6% to 7.4%. Although the absolute overhead values are lower for the classified datasets, due to their higher compressibility, the relative proportions of intra- and inter-zone redundancy closely mirror those found in the reflectance datasets.

The two-thirds to one-third split between tile and zone overlaps also closely corresponds to the 20% and 13% additional land pixels identified in the spatial overhead experiments. This strong alignment between the spatial and volumetric analyses not only validates the methodological approach applied in this thesis but also reinforces the interpretation of the results. It clearly highlights the structural inefficiencies inherent to the S-2 TGS, in which both intra- and inter-zone overlaps contribute significantly to the overall data volume inflation.

Moreover, it becomes increasingly clear that any meaningful optimisation of the system must address both sources of redundancy. Focusing on only one would not lead to sufficiently satisfactory results. This is particularly evident in the fact that, even after eliminating the S-2 intra-zone overlaps, the data volumes of the E7 versions remain approximately 10% lower in the case of the reflectance datasets, for example. This further highlights the advantage of

systems like the E7G, which inherently avoid unnecessary zone-induced redundancy through their design and thereby achieve a more efficient and compact spatial structure.

5 Conclusion and outlook

Summary and implications

The redundancy inherent in the S-2 TGS has previously been quantified through a comparison of the total mapped land area to the actual terrestrial surface. While these experiments provided valuable estimates of system-induced redundancy in terms of spatial overhead, they were limited to binary land/non-land data and did not consider actual imagery. This thesis builds upon those prior investigations by transferring the same methodological approach to real datasets. A volume-based metric, the data volume overhead, was introduced as a practical equivalent to spatial overhead, allowing for the quantification of storage inefficiencies caused by the structural design of the TGS. By comparing the volumes of datasets stored in the S-2 TGS to reference volumes that are nearly free of redundancy, the analysis assesses the impact of the TGS design on actual storage demands.

The global analysis of the non-polar land domain demonstrated that the S-2 TGS causes substantial data volume overheads due to its structural design. It can be observed that the amounts of redundant and thus unnecessarily stored data vary depending on the data type. Reflectance datasets, due to the continuous nature of their pixel values, exhibit higher redundancy, with overheads of about 32%, whereas classification datasets, characterised by discrete value ranges and lower variance, show slightly lower values of around 29%. Varying the spatial sampling, in contrast, led to nearly identical data volume overheads and thus seems to have no significant influence. However, this conclusion currently rests only on the reflectance datasets. Overall, the observed data volume overheads in this global setup confirm the order of magnitude established in the earlier spatial overhead experiments and suggest that significantly more data volume is stored within the S-2 TGS than would be technically necessary.

The findings from the global experiment also hold true for the geographically constrained setups of individual continents. Additionally, a latitude-dependent effect already identified in the spatial overhead analyses was confirmed. Due to the UTM system setup, higher geographic latitudes exhibit greater tile overlap densities. As a result, Europe and North America exhibit higher data volume overheads compared to more equatorial regions such as Africa or South America.

Since the measured data volumes refer to compressed file sizes rather than raw data, the potential influence of compression on the results was also investigated. While compression ratios vary significantly between datasets due to their content properties, they remain nearly identical across all TGS versions of the same dataset. This indicates that the observed differences in data volumes are indeed driven by the redundant spatial mapping introduced by the respective TGS, and image compression does not significantly distort the results.

The analysis of overlap types within the Sentinel-2 TGS revealed the respective contributions of intra-zone and inter-zone redundancies to the overall overhead. Approximately two-thirds of the original data volume overhead disappears when tiles within the same TG are mosaicked, indicating that most redundant storage stems from overlaps within the same UTM zones. However, still about one-third of the overhead persists, resulting from overlaps between adjacent zones. This breakdown closely matches the findings from the spatial overhead experiments and again validates the volume-based methodology employed in this thesis. Moreover, it shows that both sources of redundancy contribute substantially to total overhead. Consequently, potential changes made to the system in order to make it more efficient must address both types of overlap. The results further highlight the benefits of spatial reference systems specifically designed for efficient handling of EO raster data, such as the Equi7Grid, which inherently minimises such overlaps through globally sophisticated mapping and tiling strategies.

Future research

The methodology introduced in this thesis to quantify data volume redundancy in the S-2 TGS can be extended in multiple directions:

- For the analyses, out of the S2M data collection, the red spectral band was chosen as a representative for S-2 reflectance data. To broaden the analysis, other spectral bands, such as the blue or green channel, could be included. This would give insight into whether data of the same bit type but representing different physical properties yields different data volumes. In general, the analyses could be extended beyond optical surface reflectance and classification data. For instance, Synthetic Aperture Radar or derived products such as vegetation indices or biophysical variables may exhibit distinct volume characteristics and offer insight into how different categories of data respond to being mapped redundantly.

- The results showed that the 10m classification dataset, due to resampling, contains artificially duplicated pixels, significantly altering its compression behaviour and thereby skewing the data volume overheads. Thus, the earlier conclusion that spatial resolution is a non-significant driver of the volume overhead currently only applies to the reflectance data. Further analysis of datasets with alternative spatial resolutions, beyond 10 or 20m, would help clarify how redundancy scales with pixel size. However, any resolution used must be supported by both the S-2 and reference TGS. This is precisely why the 60m datasets of the S2M were excluded from this study, as the E7G T1 tile size does not support them. Analysing such data would therefore require an adapted reference volume estimation approach.

- Finally, the methodological framework developed in this thesis could be applied to analyse entirely different TGSs beyond S-2 since the data conversion workflow sup-

ports transformation between all systems that comply with the TGS definition criteria. Thus, a dataset stored in any candidate TGS can be converted into the E7G to further generate the required reference version, as demonstrated in this thesis. A candidate for future analysis could be the Landsat ARD Tiling Grid. However, at the time of writing, this system is not yet globally defined and is currently only nationally implemented.

The results presented in this thesis demonstrate that the choice of spatial reference system has a measurable and substantial impact on the efficiency of Earth Observation data handling. In an era of exponentially growing satellite archives and datacube infrastructures, even moderate improvements in storage efficiency can lead to considerable downstream benefits for processing, dissemination, and accessibility (CEOS, 2017). As shown here, the structural design of a TGS alone can induce data volume overheads of up to one-third. This highlights the critical importance of spatial frameworks that are not just geometrically rigorous but also operationally efficient.

Bibliography

Bauer-Marschallinger, B., & Falkner, K. (2023). Wasting petabytes: A survey of the sentinel-2 utm tiling grid and its spatial overhead. *ISPRS Journal of Photogrammetry and Remote Sensing*, *202*, 682–690. https://doi.org/10.1016/j.isprsjprs.2023.07.015

Bauer-Marschallinger, B., Sabel, D., & Wagner, W. (2014). Optimisation of global grids for high-resolution remote sensing data. *Computers & Geosciences*, *72*, 84–93. https://www.sciencedirect.com/science/article/pii/S0098300414001629

Bruce, C. A., & Ruth, M. (2019). *OGC GeoTIFF Standard Version 1.1* (tech. rep. No. 19-008r4) (OGC Implementation Standard). Open Geospatial Consortium. https://docs.ogc.org/is/19-008r4/19-008r4.html

Buchroithner, M. F., & Pfahlbusch, R. (2017). Geodetic grids in authoritative maps–new findings about the origin of the UTM Grid. *Cartography and Geographic Information Science*, *44*(3), 186–200. https://doi.org/10.1080/15230406.2015.1128851

Castriotta, A. G. (2023). *Copernicus Sentinel Data Access Annual Report 2022* (Annual Report No. COPE-SERCO-RP-23-1493) (Issue 1 Rev 0, Date: 2023-05-26). European Space Agency. European Space Operations Centre, Darmstadt, Germany.

CEOS. (2017). *The CEOS Open Data Cube Initiative* (tech. rep.) (Accessed: 2025-06-04). Committee on Earth Observation Satellites (CEOS). https://www.opendatacube.org/_files/ugd/f9d4ea_1aea90c5bb7149c8a730890c0f791496.pdf

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C., Skakun, S. V., & Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote sensing of environment*, *219*, 145–161. https://doi.org/10.1016/j.rse.2018.09.002

DMA. (1989). Defence Mapping Agency: The universal grids: Universal Transverse Mercator (UTM) and Universal Polar Stereographic (UPS). *(DMA Technical Manual 8358.2)*.

Dwyer, J. L., Roy, D. P., Sauer, B., Jenkerson, C. B., Zhang, H. K., & Lymburner, L. (2018). Analysis ready data: Enabling analysis of the landsat archive. *Remote Sensing*, *10*(9), 1363. https://doi.org/10.3390/rs10091363

ESA. (2015). *European Space Agency: Sentinel-2 User Handbook*. https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook

ESA. (2021). *Sentinel-2 Level-2A Algorithm Theoretical Basis Document* (ATBD No. S2-PDGS-MPC-ATBD-L2A V2.10) (Issue 2.10, Date: 2021-11-15). European Space Agency.

ESA. (2023). *Sentinel-2 Level-1 Algorithm Theoretical Basis Document* (ATBD No. S2-PDGS-MPC-ATBD-L1 V1.1) (Issue 1.1, Date: 2023-01-31). European Space Agency.

ESA. (online). European Space Agency: Data Products: Sentinel-2 Tiling Grid KML File [Accessed on August 21st 2024]. https://sentiwiki.copernicus.eu/web/s2-products

European Commission, Joint Research Centre. (online). S2MOSAIC L2A 2017 [Accessed on 24 October 2024]. https://cidportal.jrc.ec.europa.eu/ftp/jrc-opendata/JEODPP/S2/

Finn, M. P., Steinwand, D. R., Trent, J. R., Buehler, R. A., Mattli, D. M., & Yamamoto, K. H. (2012). A program for handling map projections of small scale geospatial raster data. *Cartographic Perspectives*, (71), 53–67. https://doi.org/10.14714/CP71.61

GDAL/OGR contributors. (2025). *GDAL/OGR geospatial data abstraction software library*. Open Source Geospatial Foundation. https://doi.org/10.5281/zenodo.5884351

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

IOGP. (2022). *Understanding the EPSG Geodetic Parameter Dataset*. Version 9. London, United Kingdom.

ISO 19111:2019. (2019). *Geographic information - Referencing by coordinates* (Standard). International Organization for Standardization. Geneva, CH.

ISO 19123-1:2023. (2023). *Geographic information — Schema for coverage geometry and functions — Part 1: Fundamentals* (Standard). International Organization for Standardization. Geneva, CH.

ISO/DIS 19177-1. (2024). *Geographic information — Geospatial API for tiles — Part 1: Core* (Standard) (Draft: Edition 2024-08-01). International Organization for Standardization. Geneva, CH.

Jordahl, K., den Bossche, J. V., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., … Leblanc, F. (2020, July). *Geopandas/geopandas: V0.8.1* (Version v0.8.1). Zenodo. https://doi.org/10.5281/zenodo.3946761

Kelso, N. V., & Patterson, T. (2010). Introducing natural earth data - naturalearthdata.com. *Geographia Technica*, *5*(82-89), 25.

Kempeneers, P., & Soille, P. (2017). Optimizing sentinel-2 image selection in a big data context. *Big Earth Data*, *1*(1-2), 145–158. https://doi.org/10.1080/20964471.2017.1407489

Kimerling, A. J. (2002). Predicting data loss and duplication when resampling from equal-angle grids. *Cartography and Geographic Information Science*, *29*(2), 111–126. https://doi.org/10.1559/152304002782053297

Kopp, S., Becker, P., Doshi, A., Wright, D. J., Zhang, K., & Xu, H. (2019). Achieving the full vision of Earth Observation data cubes. *Data*, *4*(3), 94. https://doi.org/10.3390/data4030094

Lewis, A., Lacey, J., Mecklenburg, S., Ross, J., Siqueira, A., Killough, B., Szantoi, Z., Tadono, T., Rosenavist, A., Goryl, P., Miranda, N., & Hosford, S. (2018). CEOS Analysis Ready Data for Land (CARD4L) Overview. *IGARSS 2018 - 2018 IEEE International*

Geoscience and Remote Sensing Symposium, 7407–7410. https://doi.org/10.1109/IGARSS.2018.8519255

Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). Chapter 1.5: Data acquisition and digital image concepts. In *Remote sensing and image interpretation* (pp. 30–39). John Wiley & Sons.

Masek, J. G., Wulder, M. A., Markham, B., McCorkel, J., Crawford, C. J., Storey, J., & Jenstrom, D. T. (2020). Landsat 9: Empowering open science and applications through continuity. *Remote Sensing of Environment*, *248*, 111968. https://www.sciencedirect.com/science/article/pii/S0034425720303382

Mulcahy, K. (2000). Two new metrics for evaluating pixel-based change in data sets of global extent due to projection transformation. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *37*(2), 1–12. https://doi.org/10.3138/C157-258R-2202-5835

NASA GSFCb. (online). The worldwide reference system [Accessed on 12 October 2024]. https://landsat.gsfc.nasa.gov/about/the-worldwide-reference-system/

Natural Earth. (online). Natural Earth: 1:10m Physical Vectors: Land Polygons [Accessed on 24 October 2024]. https://www.naturalearthdata.com/http//www.naturalearthdata.com/download/10m/physical/ne_10m_land.zip

Richards, J. A. (2022). Correcting and registering images. In *Remote sensing digital image analysis* (Sixth edition, pp. 31–87). Springer.

Roy, D. P., Li, J., Zhang, H. K., & Yan, L. (2016). Best practices for the reprojection and resampling of sentinel-2 multi spectral instrument level 1c data. *Remote Sensing Letters*, *7*(11), 1023–1032. https://doi.org/10.1080/2150704X.2016.1212419

Sayler, K., & Glynn, T. (2022). *Landsat 9 Data Users Handbook* [Document LSDS-2082]. Version 1.0. Sioux Falls, South Dakota.

Snyder, J. P. (1987). *Map projections–A working manual* (Vol. 1395). US Government Printing Office.

TAS. (2022). *Sentinel-2 Products Specification Document* (tech. rep. No. S2-PDGS-TAS-DI-PSD) (Accessed on August 21st 2024). Thales Alenia Space. https://sentinels.copernicus.eu/web/sentinel

TUW, GitHub. (online). TU Wien: Equi7Grid: source geometries, python package, and documentation [Accessed on 03 February 2024]. https://github.com/TUW-GEO/Equi7Grid

USGS. (online). Landsat collection 2 u.s. analysis ready data [Accessed on 12 October 2024]. https://www.usgs.gov/landsat-missions/landsat-collection-2-us-analysis-ready-data

White, D. (2006). Display of pixel loss and replication in reprojecting raster data from the sinusoidal projection. *Geocarto International*, *21*(2), 19–22. doi.org/10.1080/10106040608542379

Wulder, M. A., Roy, D. P., Radeloff, V. C., Loveland, T. R., Anderson, M. C., Johnson, D. M., Healey, S., Zhu, Z., Scambos, T. A., Pahlevan, N., et al. (2022). Fifty years of landsat science and impacts. *Remote Sensing of Environment*, *280*, 113195. https://doi.org/10.1016/j.rse.2022.113195

Yan, L., Roy, D., Li, Z., Zhang, H., & Huang, H. (2018). Sentinel-2a multi-temporal misregistration characterization and an orbit-based sub-pixel registration methodology. *Remote sensing of environment*, *215*, 495–506. https://doi.org/10.1016/j.rse.2018.04.021

Zhu, Z. (2019). Science of landsat analysis ready data. *Remote Sensing*, *11*(18), 2166. https://doi.org/10.3390/rs11182166