IMPROVING DATA QUALITY AND MANAGEMENT FOR REMOTE SENSING ANALYSIS: USE-CASES AND EMERGING RESEARCH QUESTIONS

Martin Breunig [a], Paul Kuper [a]*, Friederike Reitze [a], Steven Landgraf [a]
Mulhim Al-Doori [b], Emmanuel Stefanakis [c], Hussein Abdulmuttalib[d], Zsófia Kugler [e]

[a] Geodetic Institute, Karlsruhe Institute of Technology, Germany – (martin.breunig, paul.kuper, friederike.reitze, steven.landgraf)@kit.edu
[b] University of Science & Technology of Fujairah, College of Engineering and Technology, U.A.E. – m.aldoori@ustf.ac.ae
[c]Department of Geomatics Engineering, University of Calgary, Canada – emmanuel.stefanakis@ucalgary.ca
[d]GIS Department, Dubai Municipality, U.A.E. – husseinma@dm.gov.ae
[e]Dept. of Photogramm. and Geoinformatics, Budapest Univ. of Technology and Economics, Hungary – kugler.zsofia@emk.bme.hu

ISPRS Commission IV: WG IV/4 and ICWG III/IVb

KEY WORDS: Geospatial Data Preparation, Geospatial Data Modelling, Geospatial Data Management, Geospatial Data Processing and Analysis, Big Geospatial Data, Remote Sensing Data Quality.

ABSTRACT:

During the last decades satellite remote sensing has become an emerging technology producing big data for various application fields every day. However, data quality checking as well as the long-time management of data and models are still issues to be improved. They are indispensable to guarantee smooth data integration and the reproducibility of data analysis such as carried out by machine learning models. In this paper we clarify the emerging need of improving data quality and the management of data and models in a geospatial database management system before and during data analysis. In different use cases various processes of data preparation and quality checking, integration of data across different scales and references systems, efficient data and model management, and advanced data analysis are presented in detail. Motivated by these use cases we then discuss emerging research questions concerning data preparation and data quality checking, data management, model management and data integration. Finally conclusions drawn from the paper are presented and an outlook on future research work is given.

1. INTRODUCTION

During the last decades satellite remote sensing has become an emerging technology producing big data every day rasing the question of appropriate data preparation, data quality checking and efficient data management. NASA® and COPERNICUS® platforms, for example, are highly attractive for a multitude of possible users in science as well as in the public and in the private sector. Access to the data is provided by means of data and information access services, which provide basic functionalities to download the data and to process them to some degree. However, as experience shows, typical data preparation processes still consist of many single steps and advanced skills in data handling are needed to manually extract data of a given region in parallel for subsequent scenes or to extract data of different regions in parallel for the same scene. Therefore, data preparation - as an important step before writing the data into the database - still is expensive in operator time and hinders a fast exploitation of the data. That is why tailored geospatial database operations to support data preparation should be provided to achieve efficient data handling and data analysis. During data preparation, also the quality of the data has to be checked covering different aspects of data quality. This means e.g. that errors in the data have to be detected and corrected and the reliability of the data has to be documented by the authors. Also accuracy dimensions should be considered. Furthermore, it is important to provide an efficient long term management of big remote sensing data and their corresponding analysis models in a geospatial database management system. Thus data analysis and data management should be closer integrated so that the analysis models have direct access to the operations of the geospatial database management system and integration of various data sources is supported adequately.

The paper is structured as follows: In section 2, we refer to related work followed by section 3, describing use cases for data preparation to improve data quality, integration of elevation data across various scales and reference systems, data management and model integration for data analysis as well as advanced neural network based data analysis. In section 4, emerging research questions are derived from the use cases to improve the preparation, quality checking, management and integration of remote sensing data and models. Finally, section 5 presents the conclusions drawn from the paper and gives an outlook on future research.

2. RELATED WORK

In the context of big data analysis and geospatial data management the improvement of data-driven workflows has been extensively discussed (Laney, 2001; Chen et al., 2014; Cheng et al., 2014; Lee and Kang, 2015; Breunig et al., 2016; Li et al., 2016; Werner and Chiang, 2021). In particular, parallel query support (Hahn et al., 2002) based on parallel hardware and software architectures (Xiaoqiang and Yuejin, 2010; Taylor, 2010; Lenka et al., 2017; SpatialHadoop, 2023) has been investigated. Intensive research has also been carried out in the

* Corresponding author

field of raster databases focusing on the efficient storage of raster data (Baumann et al., 1997) and services (Baumann, 2010) to improve the access on raster data and operations (Zhong et al., 2011; Ouyang et al., 2013; Hu et al., 2018). The appropriateness of existing database management systems to handle geospatial big data, has been examined by (Amirian et al., 2014; Mazroob et al., 2020) and other authors. A "tailored approach" to manage raster data, considering heterogeneous data models, was introduced by (Baumann et al., 2016). (Baumann et al., 2018) have proven specialized data cubes as a suitable concept to provide raster data interfaces for spatial and temporal data analysis. Here the code is "shipped to the data" to minimize the communication costs when transporting the data from one tool to another. As an example of a scalable geospatial data analytics cloud platform, the "Physical Analytics Integrated Repository and Services" (PAIRS) homogenize archived and real-time spatial data (Klein et al., 2015). This approach is empowered by Hadoop® holding a parallelized structure based on MapReduce (Klein et al., 2015). Parallel system architectures such as Hadoop® and Spark® distribute the computation actions to a computer cluster. They work on the basis of the Map-Reduce model (Dean and Ghemawat, 2008), which automatically distributes (Map) the calculation steps to the existing computers to execute there and merge (Reduce) the intermediate results of the map step into a solution. Concerning data analysis in remote sensing, Artificial Intelligence is a pregnant technology to support data handling (Lary, 2010; Lary et al. 2016; Mathieu and Aubrecht, 2018). Supervised or unsupervised machine learning algorithms, especially neural networks (NNs), have been frequently used for regression and classification (Bishop, 1995), image recognition and object detection (LeCun et al., 2015). Multiple radar applications, ground- and satellite-based have been proven to work with neural networks (NNs) (Qin et al., 2004; Lombacher et al., 2016). Zhu et al. use machine learning methods to develop algorithms from signal processing and Artificial Intelligence to improve the extraction of geospatial information from satellite data (Zhu et al., 2017).
However, until the present time, the data preparation and quality checking, data selection, integration and analysis of satellite data for scientific use are very time-consuming processes. Further research is necessary to support data scientists and experts from various disciplines adequately.

3. USE CASES

The following use-cases show examples how to improve the data processing workflow during data quality checking, data and model management, data integration, and data analysis in remote sensing scenarios.

3.1 Data preparation and data quality: Cleansing of Sentinel-1 SAR Data

Correctly unwrapping interferograms in Interferometric SAR (InSAR) approaches still poses a challenge due to its ill-defined nature and is still present in Differential InSAR (DInSAR) data used for time series analysis (Yu et al., 2019). Hence, the approach presented here is concerned with the automation of finding and mending phase unwrapping errors in Sentinel-1 data. In contrast to previous works, the problem is approached in a data-driven manner using semi-supervised active learning methods. In the proposed workflow, two distinct model types are trained, namely detection and correction model. The former is trained to detect an erroneous time series, and the latter is trained to suggest corrections on time steps to a human observer. The human observer inspects the proposed corrections using a graphical user interface and adapts them if needed. These newly

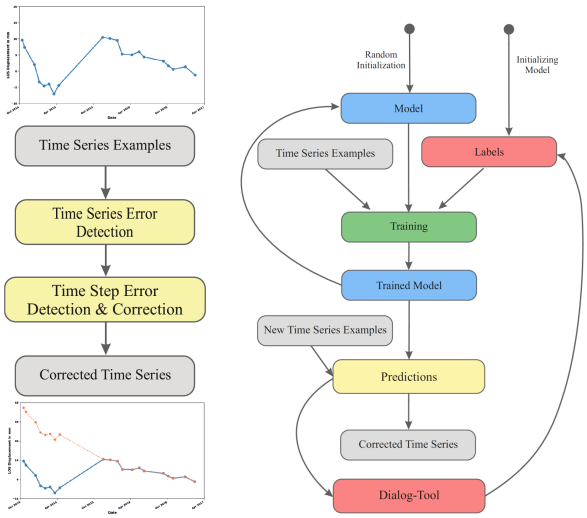

Figure 1: Workflow of error detection and correction, general overview with correction of example time series (left), and interaction loop with the user (right).

labeled time series are used for further training of the models (see Figure 1). As the provided DInSAR data is spatially sparse, i. e. not structured in a dense raster, and provides sequential structure in time, Long Short-Term Memory (LSTM) neural networks are

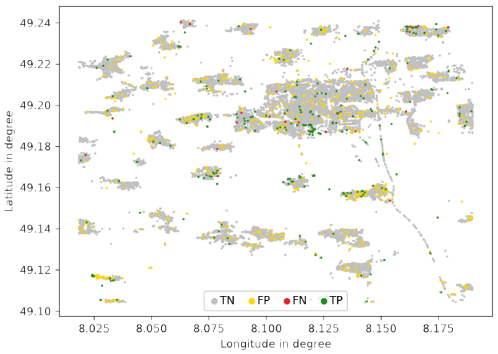

Figure 2: Example of predictions and evaluation of detection model on Landau evaluation site. TN, FP, FN, TP pixels.

used to compensate processing errors in Sentinel-1 SAR data sets of the Upper Rhine Graben (Oberrheingraben, ORG) and Landau region, Germany.

The considered data were provided by (Heck, 2019) and include two sets of processed Sentinel-1 time series. Singel Look Complex (SLC) SAR data products of the Sentinel-1 mission of the European Space Agency (ESA, 2023) were preprocessed and turned into interferograms using SNAP (Leskovec and Sosič, 2016). Afterward, Persistent Scatterer DInSAR processing was conducted using the StaMPS software (Hooper et al., 2012) with the 3D phase unwrapping technique introduced by (Hooper et al. 2007) and (Hooper, 2010).
Four training iterations were conducted (see Figure 1): First, the models were trained to imitate a simple, heuristic initializing model. Second, the models' current error detections in the Landau data set were adapted by a human observer and used as new labeled data for further training. In the last two iterations, the models were trained on labeled data of the ORG data set, first in Landau region and then on four subsets distributed over the ORG data set, where each iteration's labeled data include current predictions adapted by a human observer. All training iterations were evaluated on validation and test sets separated from the

training data regarding the true positives, false positives, false negatives, and true negatives as well as the derived metrics precision, recall, F_1-score, and balanced accuracy. One example of the trained detection model prediction on the Landau evaluation site can be seen in Figure 2. A final evaluation regarding those metrics was conducted on three evaluation sites: Landau, Staufen i. B. and Lahr, for which labels were created manually beforehand.

Regarding the detection model precision, recall, F_1-score, and balanced accuracy ranged from 0.08 to 0.28, 0.90 to 1.0, 0.14 to 0.42, and 0.93 to 0.99 respectively over the evaluation sites.

Regarding the correction model, the average metrics over all correction classes for precision ranged from 0.46 to 0.66, the recall from 0.79 to 0.92, the F_1-score from 0.4 to 0.59 and balanced accuracy from 0.72 to 0.82.

The results show that phase unwrapping errors can be detected with high recall and low to moderate precision, reducing the number of pixels a human observer must inspect. Few pixel time series are affected by PU errors and even fewer are affected in more than three time steps. Errors affecting the slope of the displacement time series are scarce and mainly found near the geothermal site at Landau where higher deformation occurs. The detection model can reduce the amount of time series required to be regarded to 3 % to 9 % of the total amount of pixels in a data set, depending on the considered area. This provides an improvement in working efficiency, as currently PU errors in PS-DInSAR data are mainly found by carefully inspecting the whole data set.

The correction is more challenging: It provides a high number of falsely corrected time steps, but shows good results in case of truly erroneous pixel time series. The portability between the two data sets is limited, but considering the highly different provided data sets, it is possible that the trained models are still able to extend to more similar data sets. Furthermore, new training iterations with training samples from another data set can be used to transfer the model to another site. In addition, different models can be stored and trained individually for different area types. In the current state the models cannot provide an automatized correction of DInSAR time series. Nonetheless, this work is a step towards the automation of error prediction and correction in geospatial data.

3.2 Integration of elevation data across various scales and reference systems

The availability of multiple diverse elevation datasets – with different coverage, vertical datum, horizontal datum and resolution, accuracies, and consideration of waterbodies is very problematic for end-users who are often required to choose amongst multiple elevation datasets for a study area. This task usually entails heavy data pre-processing, while it can lead to biased and/or inconsistent decisions, with a negative impact to the corresponding project outcomes.

In Canada, terrain datasets released by Natural Resources Canada (NRCan) mainly include the Canadian Digital Elevation Model (CDEM; NRCan, 2013) and the High Resolution Digital Elevation Model (HRDEM; NRCan, 2019). The CDEM collection is part of the NRCan's altimetry system. The coverage and resolution of the CDEM mosaic varies according to latitude and extent of the study area. With reference to the NAD83 CSRS datum, the mosaic covers the whole country at resolutions that range from 0.75 to 12 arcsec along the latitudes. The elevation values are expressed in integer meters referenced to the Canadian Geodetic Vertical Datum of 1928 (CGVD28) and can be either ground or reflective surface elevations. As a part of the CanElevation Series, the HRDEM largely improves the accuracy and spatial resolution of Canadian terrain data. The HRDEM collection consists of high-resolution DEMs derived from LiDAR and remote sensing imagery produced by separate projects. HRDEM includes a Digital Terrain Model (DTM), a Digital Surface Model (DSM) and other derived data (e.g., slope, aspect, shaded relief, color-relief, and color-shaded relief maps). HRDEM is available only over the corresponding projects footprints. In the southern part of the country, HRDEM collections include DTM and DSM datasets at a 1m or 2m resolution and projected to the UTM NAD83 (CSRS) coordinate system and the corresponding zones. In the northern part of the country, HRDEM collections include DSM datasets at a 2m resolution projected in the Polar Stereographic North coordinate system referenced to WGS84 horizontal datum or UTM NAD83 (CSRS) coordinate system. HRDEM elevation values are referenced to the Canadian Geodetic Vertical Datum of 2013 (CGVD2013), which is now the reference standard for heights across Canada.

In Li *et al.* (2021) we have recently explored the adoption of Discrete Global Grid Systems (DGGS; OGC, 2017) as an integration platform for Canadian terrain datasets to improve the coverage and elevation data quality in various study areas across the nation (Figure 3). Various algorithms were introduced to integrate the CDEM and HRDEM by direct quantization at various granularities and to aggregate the modelled elevations by mean, maximum, and minimum statistics across the resolution levels to meet the needs of different applications. For example, the minimum elevation helps to determine stream channel areas, while the maximum elevation is useful for calculating the height of vertical obstructions (Danielson and Gesch, 2011). This study set the stage for a national elevation service across various scales for Canada.

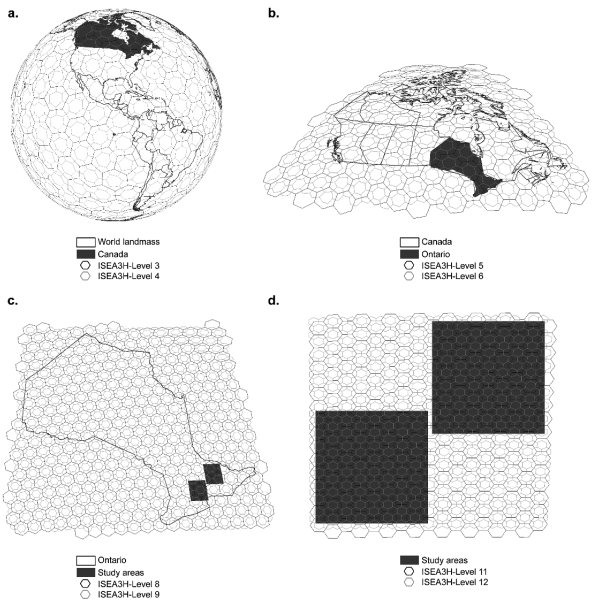

Figure 3: Representation of a. world landmass, b. Canada, c. Ontario, and d. study areas in the ISEA3H DGGS (adopted by Li *et al.,* 2021*).*

3.3 Data management and model integration for CNN-based image analysis

In the following use case precipitation data of Hyderabad, India, from 1991 – 2011 is used as Landsat 5 multi-spectral data. For categorization self-organizing neural networks (Babu, 1997), and especially Convolutional Neural Networks (CNNs) are used as

an appropriate method for image classification of satellite data (Kanellopoulos and Wilkinson, 1997). To produce a straight-forward joint workflow of data analysis and data management, and to avoid unnecessary data loading, in the following use case it is outlined how CNN-based data analysis and raster-based data and model management can work together using TensorFlow® data analysis tool and rasdaman® array-DBMS, respectively. The dataset consists of 94 images of 7 spectral bands of 1710 x 1750 with a resolution of 30m from the years 1991 to 2011 (excluding 2002-2003). Using rasdaman´s filter functions for irregular time intervals, all data have been considered presenting less than 10 percent clouds in each image (Yang, 2022). Using rasdaman, spatio-temporal database queries have been used in two different ways: first, attributes of a coverage such as bands of the 7-band hyperspectral Landsat 5 dataset have been selected via OGC´s Web Coverage Service. Secondly, OGC´s Web Coverage Processing Service have been used to execute raster operations such as special filter operations. For example, a 3D data cube with the three dimensions “band”, “time step” and “2D region” has been created as a result of a spatio-temporal database query (see figure 4). Here bands 2-7 are selected in the region of Hyderabad from 1991 to 2011, presenting one time step in figure 4.

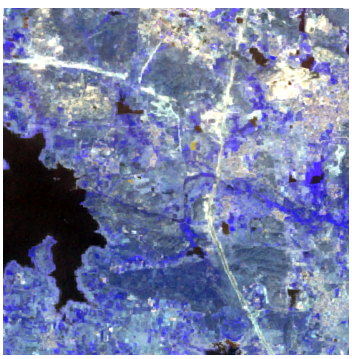

Figure 4: Result of a spatio-temporal database query represented in the three dimensions “band”, “time step” and “2D region” of a 3D data cube (from: Yang, 2022).

A database query filter operation has been used to generate False Color Composite (FCC) images facilitating the detection of vegetation changes (Yang, 2022). The user controls such queries via rasql, rasdaman´s raster query language. An example of such a database result showing two FCC images is presented in figure 5.

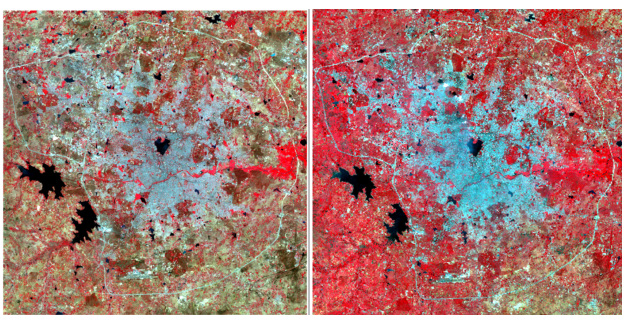

Figure 5: Result of a spatio-temporal database query computing False Color Composite images of the precipitation data in the region of Hyderabad, March and October 2011 (from: Yang, 2022).

To achieve reproducible CNN models during data analysis, the different variants of CNNs produced during the training process of the data have to be stored together with the data in the geospatial database for long term use. From a technical point of view, a direct connection is established between the data analysis tool (such as TensorFlow®) and the array-based DBMS (such as rasdaman®). This means that TensorFlow® then has direct access to the data and models stored in rasdaman®. As an alternative, graph-based DBMS may be used to store the CNNs. In both cases, a DBMS-supported image analysis workflow then consists of the following steps: Data preparation and quality checking => Data management and integration => Model management and integration => Data analysis with continuous access to data and model management. This workflow not only saves time, but supports the continuous management of data and CNN models during data analysis.

3.4 Advanced data analysis: CNN-based semantic segmentation of swimming pools

As application of advanced data analysis for remote sensing data we refer to the identification and delineation of man-made structures in aerial or satellite imagery. In this use case, aerial images from Baden-Württemberg, Germany, have been used for semantic segmentation of swimming pools and subsequent comparison with publicly available information from OpenStreetMap.
In general, such an application could enable the estimation of residential property values or monitoring of swimming pool maintenance and safety. However, for the development of robust deep learning models, a variety of requirements and challenges must be addressed:

- Large image sizes: Images need to be resized, cropped, or tiled to be usable by regular deep learning models. In this case, the images were 5000 x 5000 pixels.
- Small objects: In this case, a typical swimming pool mask has the size of approx. 100-500 pixels within a total pixel quantity of 25,000,000 pixels per image.
- Unbalanced dataset: Even after filtering out images without any swimming pools, the vast number of pixels still do not belong to the desired class.
- Geo-referencing: If the results of the semantic segmentation are supposed to be used for an application where the geospatial context is needed, the geo-referencing of the original images must be preserved.

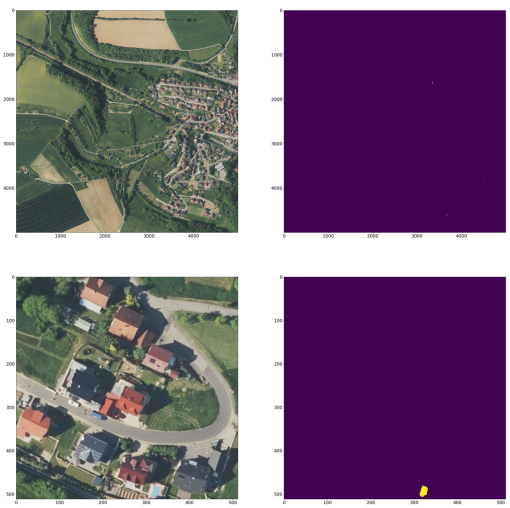

Figure 6: Top left: Original image. Top: right: Corresponding ground truth mask. Bottom left: Random crop of original image. Bottom right: Corresponding ground truth mask.

As a first step, a dataset of 155 aerial images with manually created ground truth masks was developed. As shown in Figure 6, the masks classify each pixel as either belonging to a swimming pool or background pixels. This enables the training of a deep learning model for binary segmentation to detect swimming pools in aerial images. At the same time, the coordinates of the corner points of the image are saved to preserve the geo-reference.
As a second step, a U-Net (Ronneberger et al., 2015) was trained using binary cross entropy loss and SGD optimizer (Robbins and Monro, 1951). The U-Net architecture consists of a contracting path (encoder) and an expanding path (decoder) with skip connections between them. The model was trained to perform binary segmentation on the before-mentioned dataset, where each pixel in the image is classified as either swimming pool or background. The binary cross entropy loss function measures the difference distance between the prediction and the ground truth, whereas the SGD optimizer minimizes the loss function during training by adjusting the weights of the U-Net model.

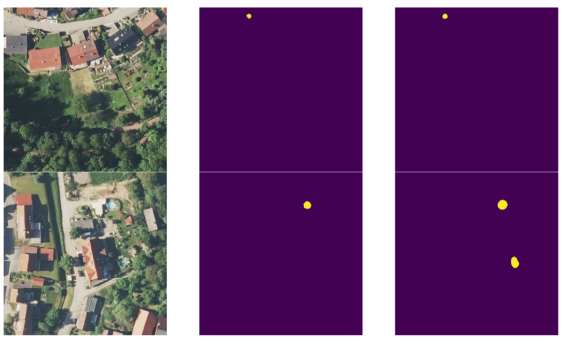

Figure 7: a) original image, b) ground truth mask, c) prediction - Preliminary segmentation results of the trained U-Net on new aerial images.

Lastly, the trained model can be used for binary segmentation of swimming pools in new aerial images. Figure 7 shows two examples of preliminary results. The first example demonstrates the effectiveness of the trained U-Net model for identifying swimming pools in aerial images as it successfully segmented the swimming pool (highlighted in yellow) from the surrounding are with high accuracy. However, the second example highlights the difficulty of the task as the model misclassified a pond as a swimming pool. This misclassification could be due to the similar shape and color of the pond and swimming pools, as well as the presence of another swimming pool and residential buildings near the pond.
The segmentation results can finally be aligned with the geo-referenced original images which enables the comparison with publicly available information from OpenStreetMap. In future work, we plan to significantly improve the segmentation results through incorporating more training data and refining the training process to thereafter identify missing swimming pools in OpenStreetMap. By doing so, we might be able to contribute to the improvement publicly available geospatial data in general.

4. EMERGING RESEARCH QUESTIONS DERIVED FROM THE USE CASES

The use cases have shown that there is a big need to improve the workflows during data preparation, data quality checking, data and model integration, and data analysis in remote sensing scenarios. We now derive general research questions triggered by the use cases.

4.1 Improving data preparation and data quality checking

To improve data preparation and data quality checking of big data, first geospatial data cleansing should be applied. Adjusting the workflow of Nelder and Wedderburn (1972) to remote sensing scenarios and following Mazroob et al. (2020), geospatial data cleansing should retain the following rules:

- Remove unwanted observations as irrelevant data: Outliers can negatively distort data models, in particular linear regression models in comparison with decision trees. Therefore, removing outliers will help model performance. Irrelevant data usually includes duplicate records, missing or incorrect information and poorly formatted data sets.
- Predict missing values - categorical or numerical, because data analysis algorithms mainly do not accept missing data: To manage missing data for categorical features, a class is added and this handles the case of no missing values. As for missing numeric data, the observation should be indicated and replaced with a “0” to satisfy the model’s algorithm requirement of no missing values enabling it to predict the best estimate for missing values rather than just the mean (Lee and Nelder, 2002).
- Remove unwanted data including duplicate, redundant and irrelevant data.

The correctness of acquired remote sensing data has to be measured by different data quality metrics. Depending on the application of the data, various quality measures can be defined. However, most important is accuracy of the remote sensing data. Accuracy defines how close the measured values are to the true values. Applied to remote sensing data, accuracy can be measured in geometric, radiometric or temporal context.
The remote sensing data lifecycle has strong relation to data quality dimensions and their adequate metrics. The definition of accuracy dimensions related to data preparation are as follows:

- *Geometric precision:* instability of the observation.
- *Spatial precision:* correctness of the spatial representation of the feature.
- *Radiometric precision/ stability:* correctness of the quantization.
- *Spectral precision:* correctness of the boundaries of the spectral bands
- *Temporal precision:* goodness of the data capture date and time.
- *Spatial accuracy:* accuracy of position of features in relation to Earth.
- *Radiometric accuracy:* correctness of the intensity values (radiance uncertainty).
- *Spectral accuracy:* correctness of the sensor’s imaging capability in the given channel.
- *Temporal accuracy/validity:* quality of the remote sensing product in time (how long does it store good information).

Besides accuracy, completeness, redundancy, readability, accessibility, and consistency are notable data quality dimensions in remote sensing.
Data quality requirements are based on data user specification. Inspection is carried out to evaluate whether data meets specification. Certain ISO standards exist for data quality check to examine fit for their intended purpose: Smart city operations are considered as a major consumer of satellite data and small-scale imagery, that is beside using high-resolution images to detail all cadastral level information and perform data

management. In fact, those smart city operations tend to insure the completeness of city smartness, e.g. governance; safety security; mobility; sustainability; livability; resilience, etc. Similar to tabular and vector-based data, raster-based remotely sensed data and its processed information including the formulated knowledge resulting from analysis and solution scenarios, that are acquired via satellite sensors/images play a major role in feeding those smart city operations. It is hence certain that raster or vector (derived from raster) data preparation, data quality checks, model integration, and data analysis directly or indirectly affect the smartness of cities, e.g. producing cloud-free and haze-free satellite images leads to more proper image interpretations and classifications allowing by that to feed operations such as managing forests in a smarter and safer way. Thus, automating actions such as accelerating the proper selection of the most effective image is an important data preparation and quality checking-based activity that in future can be performed at the level of database operations, reducing by that the expenditure in skilled operator time.

In fact, it would fare to mention that international open satellite imagery providers, as well as commercial ones, have been working extensively to ease access and encourage the usage of provided data and information, by partially performing some of the data preparation and data quality checking that require skilled labour. In other words, they not only provide corrected images based on user selection criteria but also produce the so-called Analysis Ready Data (ARD) which overcome many initial steps of data preparation and quality checking. The preparation acts for the satellite data providers can include the following among others:

- Radiometric corrections can include calibration insuring the consistency between sensors and overtime, and also atmospheric corrections of images.
- Geometric corrections such a geo-referencing, ortho-rectifications, colour balancing, mosaicking, etc.
- Metadata – including usable data masks (UDM)
- Time series of the same region.

Further, perhaps some unsupervised pre-classifications and initial segmentation can also be categorized under preparation, the parameters of which can be predicted based on provided usage and application area information. Organizations such as Copernicus (open access Hub) and Nasa, also other platforms such as ARCGIS online from ESRI, Google Earth Engine, Sen2Cube from Austria, Earth Observation browser, Earth Observation Compass, USGS, CEOS, and many others even take it to a further level, where some provide Analysis ready Data (ARD). In fact, some organizations even construct the so-called data cubes, providing users ready prepared blocks that can be combined with pre-defined tools to prepare many ready maps that are based on basic analysis functionalities, workflows and formulas. ARD maps cover atmospheric, marine and land different themes.

All produced products shall be subject to uncertainty, which entitle the necessity of quality checking. Other than checking the internal structure of the data such as duplications and missing values, error values etc., the data has to be validated and assured to reflect the respective existing reality, and that can be done by ground truth or by comparing the data to more trusted solid data sets.

4.2 Improving data and model management for CNN-based data analysis

Hitherto data analysts have to pass through a long and time-consuming process chain across several software systems to spatially or/and temporally select particular regions and time intervals out of big satellite image data. To automate and shorten the process of data selection, a geospatial database should support this process providing spatial, temporal, and spatio-temporal operations on satellite image data such as (see also Mazroob et al., 2020):

- Automatically checking geometric, topological, and temporal constraints on satellite image data to detect data errors.
- Seamlessly selecting arbitrary, a-priori not defined tiles from one given scene (defining one time step) of a satellite image.
- Selecting the same tile at different scenes (versions of the tile).
- Detecting the differences of values in pixel attributes of two scenes (change detection).
- Overlaying data from different sources and domains for the same region (e.g. SENTINEL and weather data).

“Seamlessly selecting a tile” means that the data have to be selected spatially independent of à priori fixed partitions. Furthermore, the temporal selection of the same region within a time interval has to be supported by a spatio-temporal database operation. The same is true to compute the differences between two images of the same region generated at two different time steps. Note that the overlay of two images from different data sources has to be applied carefully: the generation of “integrated models” is a sophisticated task and has to consider a variety of geometric, topological, temporal, and semantical constraints. Picking up the example of our first use case (section 3.1), the automatic checking of phase errors in interferometric synthetic aperture (InSAR) radar data could be executed by setting data constraints such as “the phase must not be greater than 2π” up to complex algorithms implementing unwrapping operations. The advantage of this database-supported approach is that the data has no more to be loaded from a data platform and later laboriously spatially or/and temporally selected as wanted, but the selection of the data is immediately done in the geospatial data management system.

Besides satellite image data, also the models used for data analysis such as Convolutional Neural Networks (CNNs) should be managed in a geospatial database management system. The persistent storage and retrieval of these models will significantly improve the reproducibility of image analysis such as CNN-based image analysis. Therefore, not only the “original models” should be stored in the database, but also optimized versions of the (learning) models. To provide the models by a database is even more important, if data from different sources are involved.

4.3 Improving data integration

The rapid growth in the number of ground, airborne, and satellite sensors has led to an unprecedented increase in the amount, variety, and rate of collection of remote sensing data. The process of integrating huge volumes of heterogeneous geospatial data using traditional data models on the geographic grid is lossy, computationally expensive, and time-consuming (OGC, 2017). Recently, Discrete Global Grid Systems (DGGS) have been proposed as an alternative geospatial reference framework that

can facilitate the fusion of multi-source remote sensing data (Goodchild, 2018; OGC, 2017; Gibb, 2021).

A DGGS applies a partitioning approach to divide the Earth's surface into a group of uniform cells at various levels of resolutions. Its hierarchical structure can effectively support the sampling, storage, modeling, processing, analysis, integration, and visualization of voluminous and heterogeneous remote sensing data.

The adoption of a DGGS in modelling and management of remote sensing data has two objectives. The first objective is to facilitate the integration of heterogeneous remote sensing datasets for a study area using a DGGS. Figure 8 summarizes the DGGS parameters, which are chosen to provide an optimal hierarchical tessellation of Canada's landmass.

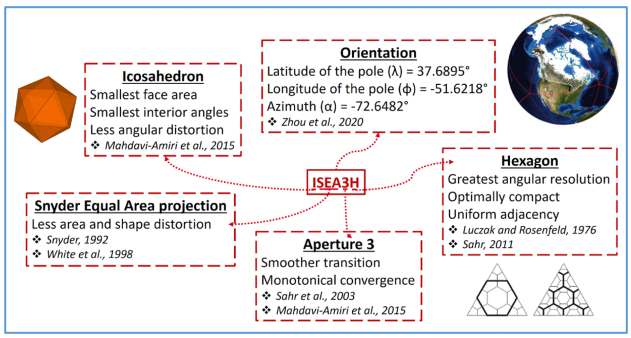

Figure 8: Example DGGS configuration parameters for Canada's landmass.

The integration of remote sensing datasets includes the following three tasks:

1. Data preparation: remote sensing data will be converted into a common horizontal and vertical datum.
2. Data quantization: remote sensing data will be assigned to DGGS cells at various resolutions based on their horizontal and vertical accuracy.
3. Quality Control: Ground control points will be used to validate the quantization of the remote sensing dataset by calculating and comparing the post-DGGS and pre-DGGS statistical errors.

The second objective is to support the geo-processing of the 'analysis-ready' remote sensing data modelled into a DGGS. Various generic and application specific analytical operations need to be developed to allow the extraction of consistent derived data from the DGGS. These operations need to be extensively validated through experimental testing for various study areas based on acquired knowledge from past works carried out by both domain experts and data analysts to guarantee an efficient toolbox for reliable decision-making processes.

5. CONCLUSIONS AND OUTLOOK

In this paper we presented use cases showing the emerging need of improving data quality and the management of data and models in a geospatial database management system before and during data analysis. Different new processes of data preparation and quality checking, integration of data across different scales and references systems as well as efficient data and model management showed that data analysis in satellite remote sensing scenarios has to be supported by various data quality and data management techniques. Based on the use cases we discussed emerging research questions concerning data preparation and data quality checking, data and model management, and data integration, cf. Figure 9.

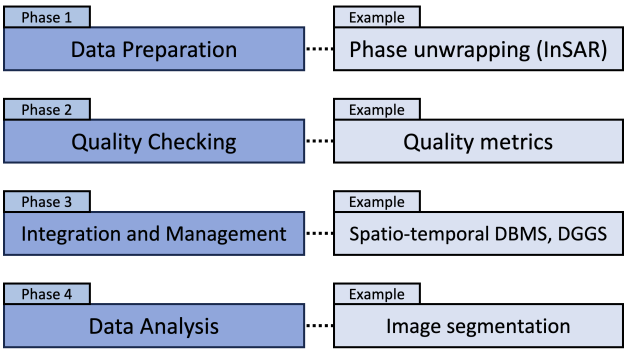

Figure 9: Overview of a typical remote sensing data processing workflow.

Data cleansing during data preparation, accuracy dimensions, the applicability on smart city operations, spatio-temporal database operations, the integration of Discrete Global Grid Systems and learning models in a geospatial database have been presented as possible solutions for an improved data analysis in remote sensing scenarios. In our future work we will embed integrated data management and data analysis architectures into existing remote sensing and geospatial applications. Finally, we intend to apply some of the introduced data preparation and data management techniques to support earth observation scenarios in the United Arab Emirates, with special emphasis on applications such as smart city management and intelligent traffic solutions.

ACKNOWLEDGEMENTS

We thank the State Office for Geoinformation and Rural Development of Baden-Württemberg, Germany for their support.

REFERENCES

Amirian P., Basiri A., Winstanley A., 2014. Evaluation of Data Management Systems for Geospatial Big Data. In: Murgante B. et al. (eds) Computational Science and Its Applications. ICCSA 2014. Lecture Notes in Computer Science, vol 8583. Springer.

Babu, G.P., 1997. Self-organizing neural networks for spatial data. Patt. Recogn. Lett. 18, pp. 133–142.

Baumann, P., 2010. The OGC Web Coverage Processing Service (WCPS) Standard, *Geoinformatica*, 14(4)2010, pp. 447-479.

Baumann, P., Furtado, P., Ritsch, R., Widmann, N., 1997. The RasDaMan Approach to Multidimensional Database Management. Proc. 12th Annual Symposium on Applied Computing (SAC'97), San Jose/USA, February 28 - March 2, 1997.

Baumann, P., Mazzetti, P., Ungar, J., Barbera, R., Barboni, D., Beccati, A., Bigagli, L., Boldrini, E., Bruno, R., Calanducci, A., Campalani, P., Clement, O., Dumitru, A., Grant, M., Herzig, P., Kakaletris, G., Laxton, J., Koltsida, P., Lipskoch, K., Mahdiraji, A.R., Mantovani, S., Merticariu, V., Messina, A., Misev, D., Natali, S., Nativi, S., Oosthoek, J., Passmore, J., Pappalardo, M., Rossi, A.P., Rundo, F., Sen, M., Sorbera, V., Sullivan, D., Torrisi, M., Trovato, L., Veratelli, M.G., Wagner, S., 2016. Big Data Analytics for Earth Sciences: the EarthServer Approach. *International Journal of Digital Earth*, 9(1), 2016, pp. 3 – 29.

Baumann, P., Misev, D., Merticariu, V., Pham Huu, B., 2018. Datacubes: Towards Space/Time Analysis-Ready Data.. In: J. Doellner, M. Jobst, P. Schmitz (eds.): Service Oriented Mapping - Changing Paradigm in Map Production and Geoinformation Management, Springer Lecture Notes in Geoinformation and Cartography. pp. 269-299.

Bishop, C., 1995. Neural Networks for Pattern Recognition, Oxford University Press, Inc. New York, NY, USA, 482p.
Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp. 5-32.

Breunig, M., Kuper, P.V., Butwilowski, E., Thomsen, A., Jahn, M., Dittrich, A., Al-Doori, M., Golovko, D., Menninghaus, M., 2016. The story of DB4GeO – A service-based geo-database architecture to support multi-dimensional data analysis and visualization. ISPRS Journal of Photogrammetry and Remote Sensing, 117, pp. 187-205.

Chen, M., Mao, S., & Liu, Y., 2014. Big data: A survey. Mobile networks and applications, 19(2), pp. 171-209. Copernicus. Open Access Hub. https://scihub.copernicus.eu/ (20 May 2019).

Cheng T., Haworth J., Anbaroglu B., Tanaksaranond G., Wang J., 2014. Spatiotemporal Data Mining. In: Fischer M., Nijkamp P. (eds) Handbook of Regional Science. Springer, Berlin, Heidelberg, pp. 1173-1193.

Danielson, J. J. and Gesch, D. B., 2011. *Global multi-resolution terrain elevation data 2010 (GMTED2010).* Reston, Virginia, USA: Earth Resources Observation and Science (EROS) Center, U.S. Geological Survey, Open-File Report 2011-1073. Available from: https://pubs.usgs.gov/of/2011/1073/pdf/of2011-1073.pdf [Accessed 1 May 2020].

Dean, J. and Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), pp.107-113.

ESA. SENTINEL-1. https://sentinel.esa.int/web/sentinel/-missions/sentinel-1 (02 March 2023).

Gibb, R. (Ed.) 2021. *Topic 21 - Discrete Global Grid Systems - Part 1 Core Reference system and Operations and Equal Area Earth Reference System*. Open Geospatial Consortium. https://docs.ogc.org/as/20-040r3/20-040r3.html

Goodchild, M. F. (2018) Reimagining the history of GIS. *Annals of GIS*, 24(1), 1–8.

Hahn, K., Reiner, B., Höfling, G., Baumann, P., 2002. Parallel Query Support for Multidimensional Data: Inter-object Parallelism. 13th International Conference on Database and Expert Systems Applications (DEXA), September 2-6, 2002, Aix en Provence, France.

Heck, A. PhD Thesis, Geodetic Institute, Faculty of Civil Engineering, Environmental and Geo-Sciences, Karlsruhe Institute of Technology (KIT), 2019.

Hooper, A., 2010. A statistical-cost approach to unwrapping the phase of InSAR time series. In Proceedings of the International Workshop on ERS SAR Interferometry, Frascati, Italy (Vol. 30).

Hooper, A., Segall, P., & Zebker, H., 2007. Persistent scatterer interferometric synthetic aperture radar for crustal deformation analysis, with application to Volcán Alcedo, Galápagos. *Journal of Geophysical Research: Solid Earth*, 112(B7).

Hooper, A., Bekaert, D., Spaans, K., and Arıkan, M., 2012. Recent advances in SAR interferometry time series analysis for measuring crustal deformation. Tectonophysics, 514-517:1–13, Jan. 2012. ISSN 00401951. doi: 10.1016/j.tecto.2011.10.013.

Hu, F., Xu, M., Yang, J., Liang, Y., Cui, K., Little, M.M., Lynnes, C.S., Duffy, D.Q. and Yang, C., 2018. Evaluating the open source data containers for handling big geospatial raster data. *ISPRS International Journal of Geo-Information*, 7(4), p.144.

Kanellopoulos, I.; Wilkinson, G.G., 1997. Strategies and best practice for neural network image classification. *Int. J. Remote Sens.*, 18, pp. 711–725.

Klein, L.J., Marianno, F.J., Albrecht, C.M., Freitag, M., Lu, S., Hinds, N., Shao, X., Rodriguez, S.B. and Hamann, H.F., 2015, October. PAIRS: A scalable geo-spatial data analytics platform. In 2015 IEEE Internat. Conference on Big Data, pp. 1290-1298.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6(70).

Lary, D.J., 2010. Artificial Intelligence in Geoscience and Remote Sensing, *Geoscience and Remote Sensing*, IntechOpen, 24p., doi: 10.5772/9104.

Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing, *Geoscience Frontiers*, 7(1), pp. 3-10, doi.org/10.1016/j.gsf.2015.07.003.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning, Nature Volume 521, pp. 436–444, doi.org/10.1038/nature14539.

Lee, J.-G., Kang, M., 2015. Geospatial Big Data: Challenges and Opportunities, Big Data Research 2 (2015), pp. 74-81.

Lenka, R. K., Barik, R. K., Gupta, N., Ali, S. M. ; Rath, A., Dubey, H., 2017. Comparative Analysis of SpatialHadoop and GeoSpark for Geospatial Big Data Analytics, 6p., Arxiv ID: 1612.07433.
Leskovec, J. and Sosič, R., 2016. SNAP: A general-purpose network analysis and graph-mining library. ACM Transactions on Intelligent Systems and Technology (TIST), 8(1):1, 2016.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., Cheng, T., 2016. Geospatial big data handling theory and methods: A review and research challenges, ISPRS Journal of Photogrammetry and Remote Sensing, May 2016, Vol.115, pp.119-133.

Li., M., McGrath, H., and Stefanakis, E., 2021 Integration of Heterogeneous Terrain Data into Discrete Global Grid Systems. Cartography and Geographic Information Science (CaGIS). Taylor& Francis. https://doi.org/10.1080/15230406.2021.1966648

Lombacher, J., Hahn, M., Dickmann, J., Wöhler, C., 2016. Potential of radar for static object classification using deep learning methods, IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM), San Diego, CA, 2016, pp. 1-4, doi: 10.1109/ICMIM.2016.7533931.

Mathieu, P.-P., Aubrecht, C., 2018. Earth Observation Open Science and Innovation, ISSI Scientific Report Series, ISBN: 9783319389677.

Mazroob, N., Breunig, M., Al-Doori, M., Heck, A., Kuper, P., Kutterer, H., 2020. Towards Intelligent Geodatabase Support for Earth System Observation: Improving the Preparation and Analysis of Big Spatio-Temporal Raster Data. Int. Arch. Photogramm. Remote Sensing. Spatial Inf. Sci., XLIII-B4-2020, Commission IV, WG IV/7, pp. 485-492, https://doi.org/10.5194/isprs-archives-XLIII-B4-2020-485-2020.

Nelder,, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. Journal of the Royal Statistical Society, Series A, 19(3), pp. 92-100.

NRCan, 2013. *Canadian Digital Elevation Model product specifications* [online]. Available from: https://ftp.maps.canada.ca/pub/nrcan_rncan/elevation/cdem_mnec/doc/CDEM_product_specs.pdf [Accessed 1 March 2021].

NRCan, 2019. *High Resolution Digital Elevation Model (HRDEM) – CanElevation Series – product specifications* [online]. Available from: https://ftp.maps.canada.ca/pub/elevation/dem_mne/highresolution_hauteresolution/HRDEM_Product_Specification.pdf [Accessed 1 March 2021].

OGC, 2017. Topic 21: discrete global grid system abstract specification [online]. Available from: http://www.opengis.net/doc/AS/dggs/1.0 [Accessed 15 November 2019].

Ouyang, L., Huang, J., Wu, X. and Yu, B., 2013. Parallel access optimization technique for geographic raster data. *Geo-Informatics in Resource Management and Sustainable Ecosystem*, pp. 533-542.

Qin, Q., Gillies, R.R., Lu, R., Chen, S., 2004. An Integration of Wavelet Analysis and Neural Networks in Synthetic Aperture Radar Image Classification, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XXXV-B2, pp. 181-186.

Robbins, H., Monro, S., 1951. A stochastic approximation method. *The annals of mathematical statistics* (1951): pp. 400-407.

Ronneberger, O., Fischer, P., Brox, Th., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer International Publishing, 2015.

Rumelhart, D.E.; Hinton, G.E.; Williams, R.J., 1986. Parallel Distributed Processing. MIT Press: Cambridge, MA, USA.

Spatial Hadoop, 2023. SpatialHadoop – a MapReduce Framework for Spatial Data, http://spatialhadoop.cs.umn.edu/ (last accessed: 02.03.2023).

Taylor, R. C., 2010. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. In BMC bioinformatics, BioMed Central.

Werner, M., Chiang, Y.-Y., Handbook of Big Geospatial Data, Springer, 2021, https://doi.org/10.1007/978-3-030-55462-0

Xiaoqiang, Y., & Yuejin, D., 2010. Exploration of cloud computing technologies for geographic information services. In Geoinformatics, 18th International Conference, pp. 1-5.

Yang, X., 2022. Data Management and Model Integration for CNN-based Image Analysis for Remote Sensing Data, Master Thesis, Geodetic Institute, Faculty for Civil Engineering, Geo- and Environmental Sciences, KIT, 73pp.

Yu, H., Lan, Y., Yuan, Z., Xu, J., and Lee, H., 2019. Phase Unwrapping in InSAR : A Review. IEEE Geoscience and Remote Sensing Magazine, 7(1):40–58, Mar. 2019. ISSN 2168-6831, 2473-2397, 2373-7468. doi: 10.1109/MGRS.2018.2873644

Zhong, Y., Sun, S., Liao, H., Zhao, Y. and Fang, J., 2011. A novel method to manage very large raster data on distributed key-value storage system. *19th International Conference on Geoinformatics* (2011), pp. 1-6.

Zhu X., Tuia D., Mou L., Xia G., Zhang L., Xu F., Fraundorfer F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources, IEEE Geoscience and Remote Sensing Magazine, 5(4), 8-36