TOWARDS INTELLIGENT GEO-DATABASE SUPPORT FOR EARTH SYSTEM OBSERVATION:

IMPROVING THE PREPARATION AND ANALYSIS OF BIG SPATIO-TEMPORAL RASTER DATA

N. Mazroob Semnani [a], M. Breunig [a], M. Al-Doori [b], A. Heck [a], P. Kuper [a]*, H. Kutterer [a],

[a] Geodetic Institute, Karlsruhe Institute of Technology, Germany - (mazroob, martin.breunig, alexandra.heck, kuper, hansjoerg.kutterer)@kit.edu

[b] College of Information Technology, University of Fujairah, United Arab Emirates – maldoori@uof.ac.ae

ISPRS Commission IV, WG IV/7

KEY WORDS: Spatio-Temporal Data Management, Spatio-Temporal Data Processing, Big Geospatial Raster Data, Intelligent Geospatial Data Analysis.

ABSTRACT:

The European COPERNICUS program provides an unprecedented breakthrough in the broad use and application of satellite remote sensing data. Maintained on a sustainable basis, the COPERNICUS system is operated on a free-and-open data policy. Its guaranteed availability in the long term attracts a broader community to remote sensing applications. In general, the increasing amount of satellite remote sensing data opens the door to the diverse and advanced analysis of this data for earth system science.
However, the preparation of the data for dedicated processing is still inefficient as it requires time-consuming operator interaction based on advanced technical skills. Thus, the involved scientists have to spend significant parts of the available project budget rather on data preparation than on science. In addition, the analysis of the rich content of the remote sensing data requires new concepts for better extraction of promising structures and signals as an effective basis for further analysis.
In this paper we propose approaches to improve the preparation of satellite remote sensing data by a geo-database. Thus the time needed and the errors possibly introduced by human interaction are minimized. In addition, it is recommended to improve data quality and the analysis of the data by incorporating Artificial Intelligence methods. A use case for data preparation and analysis is presented for earth surface deformation analysis in the Upper Rhine Valley, Germany, based on Persistent Scatterer Interferometric Synthetic Aperture Radar data. Finally, we give an outlook on our future research.

1. INTRODUCTION

During the last decades satellite remote sensing has become an important tool both in scientific earth observation and in data provision for informed decisions in politics and public administration. For this purpose, the European Commission established the COPERNICUS® programme in 2014. For the first time, a multitude of satellite remote sensing data are available - free and open - on a long-term perspective. This allows the full coverage of the earth's surface with a high temporal resolution. Using, e.g., SENTINEL-1 radar data it is both possible to derive a highly resolved digital terrain model of the earth as well as precise information about surface deformation.

The mentioned features of the COPERNICUS program 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 the Data and Information Access Services (DIAS), which provide basic functionalities to download the data and to process them to some degree. However, typical data preparation sequences consist of many single steps and correspondingly advanced skills in data handling are still needed, e.g. for the manual extraction of data for a given region in subsequent scenes. These steps are expensive in operator time and hinder a fast exploitation of the data for the respective application.

One of the main issues that we know about big spatio-temporal raster data is a lack of new tools to use available metadata without the problems mentioned before. The new concept should be efficient in terms of required run-time and effective in a way that the interesting parts of an investigation area are automatically selected. Obviously, data analytics should be improved by fitting Artificial Intelligence and supervised machine learning methods. Then characteristic phenomena can be searched across different regions and for different time steps. This refers, e.g., to the automatic selection of model components in data processing such as for the description of changes for interesting regions.

In this contribution we describe tailored geo-database operations for data preparation. Furthermore, we propose enhanced Machine Learning methods to analyse satellite remote sensing data. The paper is structured as follows: In section 2 we refer to related work followed by section 3 describing our approaches for data preparation. In section 4 a use-case is introduced focusing on the preparation and analysis of SENTINEL-1 radar data to monitor the earth surface deformation in the Upper Rhine Valley, Germany. Finally, section 5 summarizes the paper and gives an outlook on our future research.

2. RELATED WORK

In the context of big data analysis as well as 3D geo-information science the improvement of data preparation and analysis for spatio-temporal data has been extensively discussed (Breunig and Zlatanova, 2011; Chen et al., 2014; Lee and Kang, 2015; Laney, 2001; Liu et al., 2009; Li et al., 2016; Mazroob et al., 2018). In particular, parallel query support (Hahn et al., 2002)

* Corresponding author

based on parallel hardware and software architectures (Xiaoqiang and Yuejin, 2010; Sugumaran et al., 2012; Lenka et al., 2017; IBM big data and analytics hub, 2019; SpatialHadoop, 2019) has been investigated. Intensive research has also been carried out in the field of raster databases focusing on the efficient storage of raster data (Baumann et al., 1997) and services to improve the access on raster data for applications in the geosciences (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 several authors (Amirian et al., 2014; Mazroob et al., 2018). A tailored approach for raster data management in geoscientific applications considering special requirements, among other things, heterogeneous data models, has been introduced by (Baumann et al., 2015). Baumann et al (Baumann et al., 2018) have proven data cubes as a suitable concept to provide raster data for spatial and temporal data analysis, so the code "shipped to the data" is used to minimize the communication effort when transporting the data from one tool to another.

As an example of a scalable geospatial data analytics cloud platform, Physical Analytics Integrated Repository and Services (PAIRS) homogenizes archived and real-time spatial data (Klein et al., 2015). It is empowered by Hadoop® and holds a parallelized structure by MapReduce (Klein et al., 2015). With the aid of distributed file systems such as HDFS® and XtreemFS®, the data can be classified for storage and access and then the parallel system architectures such as Hadoop® and Spark® distribute the computation actions to different computers. They work on the basis of the Map-Reduce model (Dean et al., 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 (Geospatial World, 2019). The "Divide and Recombine" concept parallelizes data processing methods to significantly reduce the runtime of methods. The process begins with dividing a large amount of data into smaller subsets and with calculating the partial result for each subset in parallel. After all, these partial results will be recombined to a global result.

In the fields of geoscience and remote sensing, Artificial Intelligence is a pregnant technology to support data handling (Mathieu and Aubrecht, 2018). Supervised or unsupervised machine learning algorithms, especially neural networks (NNs), have been frequently used for regression and classification (Haykin, 1994; Bishop, 1995), image recognition and object detection (LeCun et al., 2015). Used for classification applications such algorithms are usually combined with Support Vector Machines (SVMs) (Vapnik, 1998) learning from training datasets. For geoscience and remote sensing Lary identified three application themes to use AI: code acceleration, empirical learning, and classification (Lary, 2010; Lary et al., 2016). Multiple radar applications, ground- and satellite-based have been proven to work with neural networks (NNs) (Qin et al., 2004; Alipour Fard et al., 2014; Lombacher et al., 2016). Also, geoscientific applications such as the monitoring of landslides have been supported by machine-learning methods and produced promising results (Korup and Stolle, 2014). Zhu et al. use machine learning methods to develop algorithms from signal processing and Artificial Intelligence to improve the extraction of geoinformation from satellite data (Zhu et al., 2017).

However, until the present time, the preparation and analysis of even small-scale satellite data for scientific use and data analysis are very time-consuming processes.

3. DATABASE-SUPPORTED PREPARATION AND AI-BASED ANALYSIS OF RASTER DATA

Generally, we can distinguish between two different types of applications dealing with big spatio-temporal raster data:

1) Applications dealing with very large data stores that need to be processed as a whole or at least very big parts of them at one time in a batch process. In this type of applications data are static, but database queries may change dynamically, e.g. formulated in the declarative Structured Query Language (SQL).

2) Applications dealing with very large data streams to be processed in small pieces, but in a real time or near to real time. In this type of applications database queries are static, i.e. constant during an à priori determined period of time, but the data are dynamically changing.

Examples of the first type are applications analysing the earth system a posteriori by interpreting big sets of satellite data. An example for the second type of applications is real time monitoring of extreme events such as volcanic eruptions or landslides. Because of their complexity, geospatial big data stream systems demand particular techniques and algorithms such as distributed computing and interactive analysis (Amirian et al., 2014; Beilschmidt et al., 2017).

In this paper we will concentrate on the first type of applications.

3.1 Improving data preparation

So far, geoscientists and remote sensing experts have to pass through a long process chain across several geo-software systems to spatially or/and temporally select especially interesting regions or time intervals out of big raster data (see also section 4). Furthermore, data errors are detected manually. To automate and shorten the process of data preparation significantly, a geo-database should provide spatial, temporal, and spatio-temporal operations on raster data such as:

- "Seamless cutting-out" of an arbitrary region.
- Intersecting the same region at different time steps (e.g. intersection of regions from 25 scenes).
- Determining the differences of the pixel attributes for a region between two time steps.
- Overlaying raster data from different sources and semantics for the same region (e.g. SENTINEL and weather data).
- Automatically checking geometric, topological, and temporal constraints on raster data to detect data errors.

To cut out a region seamlessly means that the data have to be selected spatially independent of stripes (into which the satellite data is divided) or other - à priori fixed - partitions of the data. E.g. queries across several stripes have to be provided simultaneously. Furthermore, the temporal selection of the same region for several time steps has to be supported by a database operation. Another important database query should compute the differences between two images of the same region generated at two different time steps. Note that the overlay of raster data from different sources has to be executed carefully considering different semantics and data models. Thus the generation of "integrated models" is a sophisticated task that has to be designed in detail considering a variety of geometric, topological, and temporal constraints. The automatic checking of phase errors in interferometric synthetic aperture (InSAR) radar data can be executed by setting data constraints such as "the phase must not

be greater than 2π" up to complex algorithms such as unwrapping algorithms.

The implementation of the mentioned geo-database operations requires compliance with additional constraints using meta data such as the associated reference system:

- The reference system of the data for each region has to be known.
- The reference systems of overlaid regions have to be equivalent to each other or to be transformed to be equivalent.
- The boundaries of cut out regions must not change within a given reference system, i.e. the boundaries must be "stable" before inserting the regions again into the geo-database after external editing.

Thus geodetic knowledge about different coordinate systems is indispensable for data integration.

3.2 Improving the data analysis

Using Geo-data cleansing
Geo-data cleansing - seen as the procedure of correcting inaccurate, redundant and corrupt geo-data - can be interpreted as the last step of data preparation as well as the first step of data analysis. Adapting the workflow of Nelder and Wedderburn (1972) we propose to execute the following steps of geo-data cleansing:

- Remove unwanted data including duplicate, redundant, and irrelevant data.
- Verify which version of the results has to be adopted.
- 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).
- Delete 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.

Using pattern recognition and classification
AI techniques utilize algorithmic models to analyse data. The presence of spatial relationships in satellite pictures is known, and a rising method for displaying these relationships is to adjust existing AI calculations demonstrated to be powerful for investigating spatio-temporal data. Of a few techniques in AI that we can use in pattern analysis for remote sensing radar data are artificial neural networks (ANNs) and kernel methods such as support vector machines (SVM), which utilize kernels to complete nonlinear regression or pattern classification (Haworth et al. 2014). Another classification method we can utilize is Random Forest (RF), because it joins various decision trees through bootstrapping (Cutler et al., 2007).

Other suitable knowledge detection (KD) undertakings to find patterns in radar data, is clustering as type of unsupervised learning. This method is preferred to reveal unclear or hidden structures in a data set and to identify hot spots (Nakaya and Yano, 2010) and anomalies in the data. This procedure involves finding irregular occasions or examples in data and it requires the definition of regular and anomalous occasions and examples, which, as the case in spatio-temporal procedures, may develop and change after some time. For the analysis of patterns in satellite data in general and InSAR data in particular, spatio-temporal clustering (STC) strategies such as ST-DBSCAN and space-time scan statistics (STSS) (Kulldorff et al., 2005) can be used to search for spatio-temporal clusters. In case of nonlinearity in spatial data, and multi-scale issue and heterogeneity (Foresti et al., 2011), KD and STDM strategies are to be used. An advantage is that the calculation in Kernel techniques such as Hadoop® kernel over extensive informational collections requires only moderately high registering resources particularly on an account of measurable properties of a space-time arrangement changing after some time. Substantial informational collections require the use of strategies conveying sets of effectively held information models to data streams. Parallel and network calculation can likewise be utilized to improve the success of pattern recognition techniques (Harris et al., 2010). Be that as it may, there are data issues that can't be enhanced with improvements of computational proficiency alone. For instance, the problem in STC is to display how clusters develop, change, move and disperse/vanish in time. This can be accomplished reflectively yet is extremely hard to evaluate in time basic applications.

Using Artificial Neural Networks
The use of artificial neural networks (ANN) in classification of remotely sensed data is utilized to perceive designs in environments patterns specifically the regulated Multilayer Perception (MLP) and the unsupervised self-Organizing Mapping (SOM) (Babu, 1997). In the classification procedure is an item occurrence division process that will profit by using single or multilayer perceptrons to survey the commitment to out-put associations (Kanellopoulos et al., 1997). MLP does a back propagation (BP) computation process utilizing a lot of covered up and yield layers (Rumelhart et al, 1986). The delta rule utilized in BP to invigorate the loads is known to be conflicting in its exhibition when managing numerous operational segments including the size and nature of the planning educational list, sort out building, getting ready parameters, and over-fitting issues which can be difficult (Cuiying et al, 2009). SOM frameworks may be well used as they were observed to be progressively steady in separating complex multivariate data (Wellar et al., 2006). This is brought about by SOM using an info layer that can get multi input sources and a multi measurements yield layer actualizing Euclidean separation to choose the triumphant hub with the nearest weight vector which can be refreshed and its neighbouring hubs amid preparing the system. This empowers the SOM system to hold the topological connections in the information, by which comparable picture characterization purposes of information are assembled as the neighbouring hubs in the aggressive layer (Chen, 1999).

To join estimations to shape a persistent time arrangement of detected information to 2006 it is important to represent the inclinations between informational collections. ANN is used to take in the mapping from one lot of estimations onto another as a capacity. These estimations are then utilized to create normal profiles. The reason for the NN mapping of remotely detected information is essentially to get familiar with the inclination as an element of area. So utilizing neural systems enables us to: Form a consistent record of information utilizing perceptions from a few space-borne instruments utilizing neural systems. Persevering predisposition between informational indexes can be taken care of by improving grouping calculations by utilizing

SVM to improve the relationship coefficient between informational indexes by permitting the AI calculations to 'right' the inclination brought about by for instance the surface sort and the presentation of other auxiliary factors that clarify the fluctuation between informational collections (Lary et al., 2009).

Using Support Vector Machines

To prepare Support vector machines (SVMs), which are known for great speculation execution, a quadratic programming (QP) process should be carried out, which is costly on memory prerequisites and preparing time for Big Data applications (Cristanini et al., 2000). The SVM choice capacity sends support vectors, which are a little piece of preparing data that is used to tackle a QP. Consequently, knowing the required example for the help vector ahead of time we can utilize a prepared set of answers for arrangement with a lot littler QP issues.

Advancements in SVM preparations have delivered an assortment of strategies for classifications of remotely sensed data, for example, choosing designs dependent on neighbourhood properties close the choice limit (Joachims, 1999), k-implies bunching to pick designs from the preparation set (Shin and Cho, 2003), a β-skeleton calculation to distinguish bolster vectors (Zhang and King, 2002), Mahalanobis separation to evaluate limit focuses (Abe and Inoue, 2001), and a subset of preparing precedents utilizing arbitrary testing in the decreased SVM (RSVM) setting (Lee and Mangasarian, 2001). Different systems were acquainted with lessen preparing set size, for example, to base the choice of a preparation of data on a factual certainty measure, and to utilize the insignificant separation from a preparation guide to the preparation instances of an alternate class as a model to choose designs close to the choice limit. A similar investigation of the outcomes acquired by the SVM classifiers prepared with information chosen by arbitrary inspecting, and by information chose on the separation from a preparation guide to the ideal isolating hyper-plane demonstrated that a size of preparing information can be essentially diminished without debasing the presentation of the subsequent SVM classifier. The correlation likewise demonstrated that arbitrary inspecting performs well with practically identical outcomes those acquired by the technique dependent on the ideal SVM yields and that consolidating the class appropriation data in the preparation set frequently improves the proficiency of the information determination strategies (Wang et al., 2006).

According to the Karush– Kuhn– Tucker (KKT) optimality conditions, the help vector decides the last isolating hyper-plane. In all actuality, the quantity of help vectors is required to be a lot littler than the absolute number of preparing precedents. It will improve the speed of SVM preparing altogether if the arrangement of help vectors is utilized for preparing, which will influence the answer for be actually equivalent to if the entire preparing set was utilized.

To recognize the help vectors, which are preparing precedents that are near choice limits, the full QP issue should be comprehended. Hence, the speed of SVM preparing will be improved without debasing the speculation execution in the event that we can locate a decent calculation strategy to locate a little arrangement of preparing information with high likelihood that it contains the ideal help vectors. The measure of the diminished preparing set can even now be bigger than the arrangement of wanted help vectors. The SVM preparing pace will be essentially improved if its size is a lot less than the extent of the all out preparing set. For a little preparing set, standard QP solvers, for example, MATLAB QP®, LOQO®, MINOS® and CPLEX®, schedules, can be used to get the arrangement. In any case, for an expansive preparing set, issues brought about by substantial memory necessities make arrangements obstinate. To stay away from this issue, various arrangements have been proposed utilizing SVM arrangement and the (KKT) conditions, for example, piecing which tackles a QP issue comprising of non-zero Lagrange multipliers αi from the last advance and a portion of the αi that damage the KKT conditions. The measure of the QP issue fluctuates however at long last equivalents the quantity of non-zero Lagrange multipliers. At the last advance, the whole arrangement of non-zero Lagrange multipliers are distinguished and the QP issue is unravelled (Huang and Lee, 2004).

Another arrangement, breaks the QP issue into a lot of littler arrangements of QP sub-issues which dependably has at any rate one precedent that abuses KKT conditions which will prompt the ideal arrangement (Vapink, 1984). Finally, sequential minimal optimization (SMO) is another technique - proposed to iteratively take care of QP sub-issues of size 2 - which can be unraveled systematically without summoning a quadratic streamlining agent. This strategy performed quicker by numerous requests of extent than the piecing technique.

Using Data Fusion

Data Fusion – seen as the processes of associating, correlating, and combining multiple resources of acquired data - may improve the quality of geo-data significantly. In remote sensing, often sensors provide multiple sources of data and require an automatic data management system to configure the sequencing, scheduling and to evaluate the reliability of the data sources. Thus the data fusion system detects the characteristic parameters of the received data as an entity and also detects the noise data caused by transmission. It then proceeds to estimate the classification of model parameters. The numerical model of the infused data incorporates the data estimations of varying sources of data of similar applications and removes redundant and conflicting observations data to optimize the system's performance. The multisensory data can then be used at all levels of the data processing system such as data Layer and decision layer (Han, 2018).

4. USE CASE: SENTINEL-1 RADAR DATA

4.1 Data description

The European Copernicus satellites produce one of the largest datasets in the world in the scale of a daily volume of nearby 20 terabytes. The evaluation of these datasets is more and more a technological obstacle for space research and technology and one of the main challenges by some organizations such as the German Aerospace Center (DLR), European Space Agency (ESA) and the European Union Satellite Centre (GISPoint, 2019).

In our use case we focus on analysis of the Earth's surface deformation based on measurement data of the Sentinel-1 satellite mission. The Sentinel-1, a Synthetic Aperture Radar (SAR) mission of the ESA Copernicus program consists of two polar-orbiting satellites (Sentinel-1A and Sentinel-1B) to gain continuous C-band radar imagery. Both satellites can map the globe together once every 6 days in wide-swath imaging mode. All Sentinel-1 SAR data have predefined product types and include Level-0, detected Level-1 and Level-2 ocean products. Raw Level-0 products commonly have a size of 1GB and Level-1 data between 1GB and 8GB per product. Sentinel Level-0 and Level-1 products are broken into 'slices' of prescribed length along a strip, because these stand-alone products are better manageable for the end-users (ESA, 2019).

In our studies, we work with Level-1 Single Look Complex (SLC) products with a size between 4GB and 5GB. Using StaMPS® (Stanford Method for PS) for Persistent Scatterer (PS) analysis, at least 12 interferograms are required (Hooper et al.,

2007), which results in datasets of at least 50GB. Because of the interferometric wide swath acquisition mode, one data slice consists of three swaths, which are themselves divided into eight to ten bursts.

4.2 Data structure

A variation of Standard Archive Format for Europe (SAFE) is applied to distribute and archive SENTINEL data, which includes binary image data and XML product metadata. In Table 1 and also Figure 1 we have a look into a SENTINEL product data structure (ESA, 2019). First issue regarding metadata to be addressed here is that SENTINEL metadata contains information about the geographical coverage of the entire scene, the number of bursts, their start of recording time, and their end of recording time. However, the metadata does not provide information about the direct relationship between burst number and geographic coverage.

File or Folder	Content
manifest.safe	General product information and characteristics of the measurement data in XML
Measurements folder	Measurement data and Image data in various binary format
Preview folder	Quicklooks in KML, PNG, HTML
Annotation folder	Product and calibration metadata in XML
Support folder	XML schemes about the format of the measurement

Table 1: A SENTINEL product folder

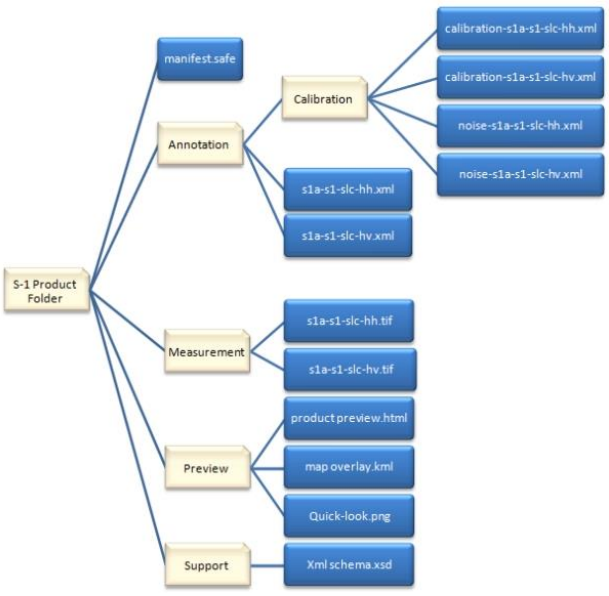

Figure 1: File Structure for Level-1 Sentinel-1 products (ESA, 2019)

4.3 Current manual data preparation

Our processing is built on the free software-packages SNAP (Sentinel Application Platform, ESA) and StaMPS (see Figure. 3). Since we work with SLC data, the data is already focused, but still exists in the slant-range geometry.

The use of the SNAP programing interface in Python enables a largely automated processing chain. However, since some steps require manual intervention, the automation is inevitably interrupted, thus increasing the time required for the processing.

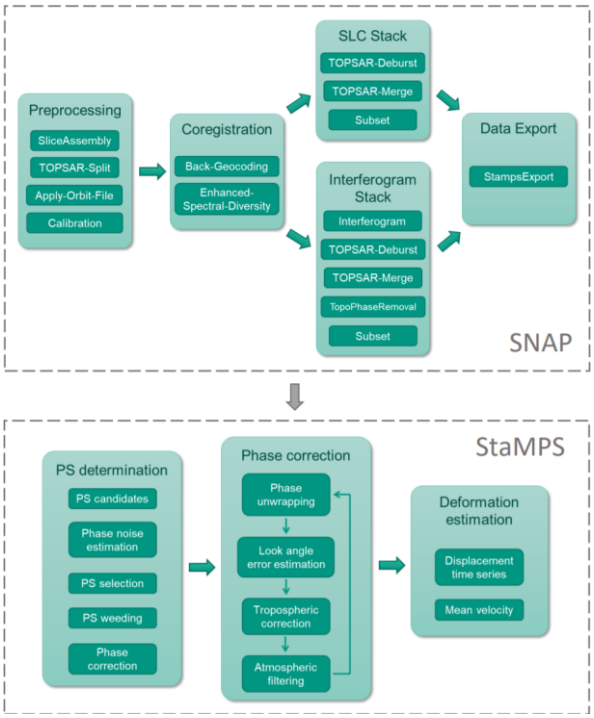

Figure 2: Workflow for PS analysis using Sentinel-1 data

The workflow begins with the viewing of the data. It must first be checked which bursts are needed to fully cover the area of interest. Each scene has to be opened separately in SNAP to determine the exact geographic coverage of the scene or individual bursts. In addition, if the region of interest is not completely covered by one slice, then two or more slices have to be joined together. If it is certain which bursts have to be selected, they can be separated using the "TOPSAR-Split" step. In our workflow the preprocessing also includes the application of the precise orbit files and a calibration of the amplitude. Since the visual aids in SNAP are limited to determining the correct bursts, a visual check of whether the correct bursts have been selected for all slices is essential after preprocessing. Without doubt this step is critical with respect to possibly undetected errors introduced by human interaction.

A visual check is also required after the "co-registration". This step needs a large amount of memory and can therefore cause individual scenes not to be correctly or completely stored without SNAP generating an error message.

The application of the step "TOPSAR-Merge" is equally depending on the location of the area of interest and the coverage of the slice. Here several swaths can be put together before the step "Subset", in which the area can be cut out. Again, it is recommended to manually check the result of these steps.

The PS analysis using StaMPS can be divided into three blocks: the determination of the PS points, phase unwrapping and the separation of the phase components containing the deformation signal, such as the estimation of the deformation.

Phase unwrapping is a crucial problem in the PS processing chain. Only the unwrapped phase relates to the Earth's surface deformation signal. However, the interferometric phase is only determined modulo 2π. The resolving of these phase ambiguities is highly prone to errors (Hanssen, R. F., 2001, Hooper, 2010).

In our workflow, we carry out visual checks after each of these blocks, adjust the parameters based on this and, if necessary, repeat the steps already taken.

Both when working with SNAP and when working with StaMPS, most of the time is spent performing visual checks and adjusting parameters based on experience. In addition, errors, e.g., due to problems with unwrapping, still occur and can only be found by visually reviewing the results.

4.4 Improving data preparation

There are some weaknesses in raster data platforms such as SNAP, which we face by a new workflow. The aim of this research is to design and develop a platform to improve the management and processing of big raster data focusing on the data from the Sentinel-1 mission.

4.4.1 Visualization of the datasets and bursts

To access a specific spatial region in the downloaded dataset in SNAP, the user cannot see the bursts in a visualized map to select them from there. Therefore, one of our missions is to visualize the data bursts on a map and prepare the possibility for the user to select the desired dataset of a desired location independent of the bursts. The user should be able to select specific datasets by individual filters such as characters, spatial locations, data sources or missions, time, applications, etc.

4.4.2 Integration of various datasets

The integration of various datasets and the ability to compare them in requested time stamps or time intervals reduces the conflicts of out-of-platform's comparison of various data sources and will increase the total speed of many user requests. The storage of heterogeneous data together needs a mutual georeferencing because of different databases, origins and imaging methods. Artificial neural network methods can be used to transform the coordinates of datasets and automatization of georeferencing process.

4.5 Improving data analysis

By the analysis of different satellite observation data, we faced heterogeneity in the data, because of the different sensor specifications of the different satellite systems and also variations in the data resolution or imaging geometry. Therefore, it is necessary to homogenize these various satellite data for the efficient and fast processing of a large amount of data (Sips et al., 2018). To homogenize the input data, with consideration of geometric reference unifier, all satellite data has to be adjusted to the specifications, which the user defines. In parallel, the accuracy for each dataset has to be calculated to guaranty the data quality. This accuracy weighting can be used for different applications to find the best data sources for data analysis. The datasets should be stored in a mature way, therefore to the partitioning of geo-data we can split them based on spatial nearness to reduce the number of items passing through a query process.

One of the obstacles to facing the analysis of big raster data is that the runtime required by an analysis method increases rapidly with the size of the input data. For example, to classify land use in low spatial resolution satellite data, the Random Forest algorithm (Breiman, 2001) needs only a few minutes to run but high spatial resolution data such as Sentinel 2 data needs weeks to classify through this algorithm. Therefore, it is necessary to adjust scalable analysis methods to handle large amounts of geodata efficiently (Sips et al., 2018). To decrease query processing time, the high frequency queried areas must be partitioned more than the less required areas.

As mentioned in section 4.3, there are some improvements needed in our workflow which can gain benefits from Artificial Intelligence methods such as Support Vector Machines, Mean-shift-clustering and neural networks. We need to provide datasets only from bursts which fully cover the queried area, otherwise the joined dataset from two or more bursts has to be generated, which exactly includes the area. The correction process and calibrations of the amplitude, the phase unwrapping errors and the estimation of the deformation has to be automized by AI methods instead of executing visual checks.

5. CONCLUSIONS AND OUTLOOK

In this paper we presented approaches to improve the preparation of spatio-temporal satellite remote sensing data by operations of a geo-database. Furthermore, proposals for the advancement of data analysis by AI methods have been presented and concreted by a use case for earth surface deformation analysis of the Upper Rhine Valley, Germany, using SENTINEL-1 raster data. In our future research we are going to overcome the difficulties in data preparation mentioned in section 4.3. We will identify typical data errors such as unwrapping errors automatically with the aid of artificial neural networks and other AI techniques. Finally, it is our goal to apply some of the introduced methods to support for earth observation applications in the United Arab Emirates.

ACKNOWLEDGEMENTS

The support of preliminary work for this research has been supported by the German Research Foundation (DFG) within BR2128/11-3.

REFERENCES

Abe, S., Inoue, T., 2001. Fast training of support vector machines by extracting boundary data. In: Proceedings of the International Conference on Artificial Neural Networks (ICANN), pp. 308–313.

Alipour Fard, T., Hasanlou, M., Arefi, H., 2014. Classifier Fusion of High-Resolution Optical and Synthetic Aperture Radar (SAR) Satellite Imagery for Classification in Urban Area, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XL-2/W3, 25-29, https://doi.org/10.5194/isprsarchives-XL-2-W3-25-2014.

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., 2015. 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.

Beilschmidt, C., Drönner, J., Mattig, M., Schmidt, M., Authmann, C., Niamir, A., Hickler, Th., Seeger, B., 2017. VAT: A Scientific Toolbox for Interactive Geodata Exploration, Datenbank-Spektrum, Vol. 17, issue 3, pp. 233-243.

Birant, D., Kut, A., 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data, *Data & Knowledge Engineering*, Volume 60, Issue 1, January 2007, pp. 208-221.

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., Zlatanova, S., 2011. 3D geo-database research: Retrospective and Future Directions, *Computers & Geosciences* 37(2001), pp. 791-803.

Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., & Han, L. D., 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Systems with Applications 36, 6164– 6173. doi: 10.1016/j.eswa.2008.07.069.

Chen, Z., 1999. Texture segmentation based on Wavelet and Kohonen network for remotely sensed images. IEEE SMC'99 Conference Proceedings, Tokyo, Japan, Oct. 12-15; Vol. 6, pp. 816–821.

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.

Cuiying, Z.; Liang, Z.; Xianyi, H. Classification of rocks surrounding tunnel based on improved BP network algorithm. Earth Sci. J. China Univ. Geosci. 2005, 30, 480–486. Remote Sens. 2009, 1264.

Cutler, D. R., Edwards, T. C. Jr., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J., 2007. Random Forests for Classification in Ecology, *Ecology*, 88(11), 2783-2792.

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 (10 May 2019).

Foresti, L., Tuia, D., Kanevski, M., & Pozdnoukhov, A., 2011. Learning wind fields with multiple kernels Stochastic Environmental Research and Risk Assessment, 25(1), pp. 51-66. doi: 10.1007/s00477-010-04050.

Geospatial World. We are entering a new age of geospatial Big Data. https://www.geospatialworld.net/article/we-are-entering-a-new-age-of-geospatial-big-data-dr-abhay-kimmatkar-ceinsys/ (10 May 2019).

GISPoint. Big Data from Space. https://www.gispoint.de/news-einzelansicht/2862-big-data-from-space-loesungen-fuer-die-datenflut-aus-dem-all-gesucht.html (10 May 2019).

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.

Han, M., 2018. Application of Artificial Intelligence Detection System Based on Multi-sensor Data Fusion, *International Journal of Online and Biomedical Engineering*, vol 14, no 26-2018.

Hanssen, R. F., 2001. Radar interferometry: data interpretation and error analysis (Vol. 2). Springer Science & Business Media.

Harris, R., Singleton, A., Grose, D., Brunsdon, C., & Longley, P., 2010. Grid- enabling Geographically Weighted Regression: A Case Study of Participation in Higher Education in England, Transactions in GIS, 14(1), pp. 43-61.

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., 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).

IBM Big Data & Analytics Hub http://www.ibmbigdatahub.com/, 2019 (last visited: 15.03.2019).

Haworth, J., Cheng, T., 2014. Graphical LASSO for local spatio-temporal neighbourhood selection. In: Proceedings the GIS Research UK 22nd Annual Conference. Presented at the GISRUK 2014, University of Glasgow, Glasgow, Scotland, pp. 425–433.

Haykin, S., 1994. Neural Networks: A Comprehensive Foundation, Prentice Hall PTR Upper Saddle River, NJ, USA.

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.

Huang, S. Y., Lee, Y. J., 2004. Reduced support vector machines: a statistical theory. Technical report, Institute of Statistical Science, Academia Sinica, Taiwan. http://www.stat.sinica.edu.tw/syhuang/.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C. J. C., Smola, A. J. (eds.): Advances in Kernel Methods - Support Vector Learning. MIT Press, Cambridge, MA, pp. 169–184.

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

Kanevski, M., Pozdnoukhov, A., and Timonin, V., 2009. Machine Learning for Spatial Environmental Data: Theory, Applications and Software, EPFL Press, 377p.

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.

Korup, O. and Stolle, A., 2014. Landslide prediction from machine learning, *Geology Today*, 30(1), pp. 26-33, doi.org/10.1111/gto.12034.

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

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., & Mostashari, F., 2005. A Space-Time Permutation Scan Statistic for Disease Outbreak Detection. PLoS Medicine, 2(3).

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.

Lee, Y. J., Mangasarian, O. L., 2001. RSVM: Reduced support vector machines. In: Proceedings of the First SIAM International Conference on Data Mining.

Lee, Y., Nelder, J.A., 2001. Modelling and analyzing correlated non-normal data. Statistical Modelling, 1, pp. 3-16.

Lee, Y. and Nelder, J.A., 2002. Analysis of ulcer data using hierarchical generalized linear models. Statistics in Medicine, 21, pp.191-202.

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.

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.

Liu, G., Zhu, Q., He, Z., Zhang, Y., Wu, C., Li, X., & Weng, Z., 2009. 3D GIS database model for efficient management of large scale underground spatial data. In Geoinformatics, 2009 17th International Conference on IEEE, pp. 1-5.

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, Semnani N., Kuper, P.V., Breunig, M., Al-Doori, M., 2018. Towards an intelligent platform for big 3D geospatial data management, ISPRS Photogramm. Remote Sens. Spatial Inf. Sci, IV-4, pp. 133-140.

Momjian. The High Value of Data. https://momjian.us/main/blogs/pgblog/2019.html#March_8_2019 (10 May 2019).

Nakaya, T., & Yano, K., 2010. Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics. *Transactions in GIS*, 14(3), 223-239. doi: 10.1111/j.1467-9671.2010.01194.x

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

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.

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

Shin, H. J., Cho, S. Z., 2003. Fast pattern selection for support vector clas-sifiers. Proceedings of the 7th Pacific-Asia Conference on Knowl-edge Discovery and Data Mining. Lecture Notes in Artificial Intelli-gence (LNAI 2637), pp. 376–387.

Sips, M., Scheffler, D., Rawald, T., Eggert, D., Hollstein, A., Segl, K. (2018): Big-Data-Ansätze für die schnelle Extraktion relevanter Informationen und Muster aus großen Datenmengen. - System Erde, 8, 1, pp. 40-45.

Spatial Hadoop, 2019. SpatialHadoop – a MapReduce Framework for Spatial Data, http://spatialhadoop.cs.umn.edu/ (last accessed: 15.03.2019).

Sugumaran, R., Burnett, J., & Blinkmann, A., 2012. Big 3d spatial data processing using cloud computing environment. In Proceedings of the 1st ACM SIGSPATIAL international workshop on analytics for big geospatial data, ACM, pp. 20-22.

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

Van Oosterom, P., Martinez-Rubi, O., Ivanova, M., Horhammer, M., Geringer, D., Ravada, S., Tijssen, T., Kodde, M. & Gonçalves, R., 2015. Massive point cloud data management: Design, implementation and execution of a point cloud benchmark, *Computers & Graphics*, 49, pp. 92-125.

Vapnik, V.N., 1998. Statistical Learning Theory, Adaptive and Learning Systems for Signal Processing, Communications and Control, John Wiley & Sons, 768p.

Wang, Y., Xue, Z., Shen, G. et al. Amino Acids, 2008. 35: 295. https://doi.org/10.1007/s00726-007-0634-9.

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

Zhang, W., King, I., 2002. Locating support vectors via β-skeleton technique. In: Proceedings of the International Conference on Neural Information Processing (ICONIP), pp.1423–1427.

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