Data Integration in Earth Sciences

Hendrik Paasche[1,12], Detlef Eberle[2], Sonali Das[3], Antony Cooper[4], Pravesh Debba[5], Peter Dietrich[1], Nontembeko Dudeni-Thlone[5], Cornelia Gläßer[6], Andrzej Kijko[7], Andreas Knobloch[8], Angela Lausch[9], Uwe Meyer[10], Edgar Stettler[11], Ulrike Werban[1]

[1] Helmholtz Centre for Environmental Research - UFZ, Department of Monitoring and Exploration Technologies, Permoserstr. 15, 04318 Leipzig, Germany

[2] Council for Geoscience, Geophysics Unit, Private Bag X112, Pretoria 0001, South Africa; now: geotec Rohstoffe GmbH, Friedrichstr. 95, 10117 Berlin, Germany

[3] Council for Scientific and Industrial Research (CSIR), Built Environment, Logistics and Quantitative Methods, P.O. Box 395, Pretoria 0001, South Africa

[4] Council for Scientific and Industrial Research (CSIR), Built Environment, Spatial Planning and Systems, P.O. Box 395, Pretoria 0001, South Africa

[5] Council for Scientific and Industrial Research (CSIR), Built Environment, Statistical Modelling and Analysis, P.O. Box 395, Pretoria 0001, South Africa

[6] Martin-Luther University Halle-Wittenberg, Institute of Geoscience and Geography, von-Seckendorff-Platz 4, 06120 Halle, Germany

[7] University of Pretoria, Aon Benfield Natural Hazard Centre, Private Bag X20, Hatfield 0028, Pretoria, South Africa

[8] Beak Consultants GmbH, Am St. Niclas Schacht 13, 09599 Freiberg, Germany

[9] Helmholtz Centre for Environmental Research - UFZ, Department of Landscape Ecology, Permoserstr. 15, 04318 Leipzig, Germany

[10] Federal Institute for Geosciences and Natural Resources, Geophysical Exploration and Technical Mineralogy, Stilleweg 2, 30655 Hannover, Germany

[11] Afrika Gold AG, Firststr. 15, 8835 Feusisberg, Switzerland; University of the Witwatersrand, School of Geoscience, Private Bag 3, Wits 2050, Johannesburg, South Africa

[12] Corresponding author: telephone: +49 341 2351414; fax: +49 341 2351939

Email: hendrik.paasche@ufz.de

Manuscript submitted for publication in *Environmental Earth Sciences, section International Views and News*, July, 2013

PREFACE

This article reflects discussions German and South African earth scientists, statisticians and risk analysts had on occasion of two workshops about Data Integration Technologies for Earth System Modelling and Resource Management. The workshops were held in October 2012 at Leipzig, Germany, and April 2013 at Pretoria, South Africa, and were attended by about 70 researchers, practitioners and data managers of both countries. Both the events enabled by financial support of the South African National Research Foundation (NRF) were arranged as part of the South African – German Year of Science 2012/13.

INTRODUCTION

Human welfare depends critically on sustainable utilization of our environment. Earth Sciences are concerned with the observation of terrestrial environments for process understanding and current state exploration. Driving forces for these efforts are manifold, e.g. the prediction of future terrestrial scenarios, particularly in response to anthropogenic impacts, ecosystem management, resource exploration, or hazard risk analysis. The results of observing and modelling terrestrial environments provide the knowledge base to handle a wide variety of societal issues, such as economy regulation or migration.

To assess the complexity of terrestrial systems adequately, a wide variety of data must be recorded (e.g. Zacharias et al. 2011). Such data may vary in spatial scale, ranging from sparse local point information, e.g. as provided by boreholes or soil samples, to satellite imagery providing spatially continuous data about the area of interest. A temporal aspect may also be present in the collected database if data are collected repeatedly or during an extended time span.

When sampling information about our environment, Earth Sciences partly overlap with neighbouring disciplines. For example, Earth scientists may be interested in assessing biologic diversity information, plant vitality, or intrusion/diminishing of species in a certain region.

Thus, the databases required to address the complexity of terrestrial environments may comprise data (i) acquired using highly different methodologies, and (ii) observed on different spatial and/or temporal scales.

Each Earth Science discipline, e.g., Geophysics, Geochemistry or Geoecology, has its own methodological expertise, which may be particularly suitable for addressing specific issues in a geophysical, mineralogical, or ecological context, respectively. Some distinct states in our environment may be satisfactorily addressed by a single sub-discipline. However, none of these disciplines provides all-embracing information about our terrestrial environment allowing for a holistic understanding of processes in terrestrial environment, e.g. regional water balance development. When striving to assess the complexity of terrestrial environment, Earth Scientists see themselves challenged by the task of analysing and integrating the information provided by various observation parameters or variables resulting in multi-method and multi-scale databases.

Computational and technological developments achieved over the last three decades resulted in an ever-increasing amount of data recorded over a distinct time period. Modern geoscientific databases covering intermediate-sized regions of several thousand square kilometres comprise billions of digital readings, e.g. in the shape of satellite imagery, geochemical and mineralogical sampling, geophysical survey data, ecological and hydrologic information, etc. The sheer size of the available amount of data and rapid data acquisition capacities are increasing the pressure on Earth Scientists to develop largely automated data analysis and integration techniques allowing for rapid information extraction from complex databases.

In Earth Sciences, the increasing computational power has been largely considered to improve the modelling of terrestrial processes, i.e., using finite elements and self-adapting meshes for highly flexible model parameterization, or simply pushing the spatial resolution limits towards higher resolution by increasing the number of model parameters (e.g. Wood et

al. 2011). However, incorporation of observed information into the modelling of processes going on in terrestrial environments is critical, particularly for “soft” data providing information that is non-linearly or even non-uniquely linked to the process of interest and the considered model parameters. For example, when modelling subsurface water flow, geophysical information, such as electrical resistivity distribution, could be used to constrain the hydrological model. However, in a deterministic sense, we do not know any quantifiable exact relationship of the physical electrical resistivity parameter with hydrologic parameters required for hydrologic process modelling, e.g. with the effective porosity, This relationship may also spatially vary because of changing sedimentary composition or evolutionary history of the deposited materials.

Nevertheless, depending on the Earth scientific discipline under consideration, a varying number of theoretical, empirical or semi-empirical deterministic transfer functions are frequently used to “convert” an observed data set into a quantity more closely linked to the model parameters. A well-known example of this kind is Archie’s law, which allows the direct conversion of electrical resistivity into porosity of granular sediments. Such deterministic transfer functions are usually calibrated using a limited number of observations where the target parameter (porosity in our Archie’s law example) and the “soft” constrain parameter (spatial electrical resistivity distribution) are commonly known. However, such immediate approach of data integration based on data conversion, particularly when going along with rather subjectively selected calibration sites , may impede high-quality modelling of terrestrial environments and may even restrict objective assessment of current states in terrestrial environments.

DATA INTEGRATION TECHNOLOGIES IN EARTH SCIENCES

Earth Sciences have a long tradition in visual integration of different or complementary data sets (e.g. Kvamme, 2006; Rink et al. 2012). Traditional exampl is the use of a light

table, which enables the semi-transparent overlay of several information layers, i.e. maps. Modern approach is to use virtual environments to establish graphically integrated composites of different data sets. Popular example is, for example, the geographical information system (GIS) . Another frequently used visual integration is the ternary plot. Here, co-located information layers are integrated by scaling red, blue and green colour saturation according to the dynamic range of each information layer. A fourth layer can be introduced when scaling the brightness (grey tone) of the composite image.

Visual integration is usually limited to a small number of information layers, e.g. no more than four in the case of a ternary plot. Visual integration requires a subsequent step of interpretation largely based on subjective insights and deductive reasoning of the human interpreter. Objective or quantitative assessment of the information extracted from a visually integrated composite image is not possible thoughdifferent interpreters may come up with analyses of the composite that are more or less concurrent . To some extent , the differences in exegesis may reflect ambiguities inherent to the the acquisition accuracy of the data. Erroneous technical understanding, subjective weighting of the importance of individual information layers or misleading of the interpreters eye by different reception of blue, green and red colours in an image may cause additional bias when integrating data visually. The use of visual data integration and information extraction prohibits consecutive objective risk analysis or reliable prediction of future scenarios. For example, when sub-dividing a survey area into a number of distinct units piecewise exhibiting similar characteristics, these sub-division may turn out to be more or less erroneous depending on the skills of the human interpreter. Applying consecutive model parameterization to the identified units cannot be expected to improve the initial spatial compartment identification based on subjective insights.

Therefore, more powerful data integration strategies strive to go beyond a simple visual integration by providing objective numerical information easing the interpretation of multiple

data sets. Geostatistical tools have been developed to analyse and describe spatial and/or spatio-temporal distributions of data where concepts of random function theory are used that consider the individual observations as well as observation gaps as correlated random variables. The uncertainty with regard to spatial estimation and simulation of data is described by a statistical model of spatial continuity, e.g. a variogram. There are statistical models generated from a specific data set which are assumed valid for the prediction of other data sets. This would present a much more quantitative way of integrating data compared to visual integration and a more flexible approach compared to utilising a deterministic transfer function, since uncertainty of data can be considered as expression of different conditioned stochastic simulation . However, meeting the conditions allowing for a common statistical description, i.e. high correlation of different data sets, usually requires the processing of the data sets whichcan only be performed to a certain limit of accuracy (option: a geostatistical example from Pravesh's presentation). Problems may also arise if data exhibit spatially or temporally highly variable correlation among data points for a given observation spacing.

Some statistical models used to describe data distributions, e.g., variograms and correlation lengths, are sometimes neither straight forward to imagine for a human brain nor do they allow for easy assessment of spatial statistical differences when analysing a varying number of observations. Instead, a human interpreter is largely looking for distinct patterns or changes in pattern of visualised data. For example, one might be interested in identifying sub-areas exhibiting similar characteristics within one or between different data sets. Conversely, one may be interested to assess the boundaries between such sub-areas. The outline of sub-areas may also pay attention to internal data characteristics, e.g., varying quality of observations reflected by different noise levels, or by evidence of distinct shapes in visualised data, e.g. lineaments representing dykes or faults, or circular features reflecting pipes and vents. Such pattern recognition in discrete databases has undergone a significant boost over the last decade stimulated by algorithmic developments mainly coming from fields of

computer science, e.g. machine learning, data mining and image processing.

Pattern analysis techniques classifying multi-feature observations, such as cluster analyses, support vector machines, or artificial neural networks, became popular for a number of applications in Earth sciences. These pattern recognition techniques are highly flexible when it comes to the analysis and integration of disparate data sets with spatial measurement variability, i.e. data sets with spatially varying degree of correlation . The individual disciplines in Earth sciences feel currently attracted by these techniques to a varying degree. Software modules integrating the multiple information layers and analysing the information content in a probabilistic sense using above techniques have been developed for specific tasks, such as the prediction of the mineral potential of a specific area. (option: image from Presentation of Andreas Barth). Data integration techniques which provide a probabilistic formulation of the results are based on Bayesian inference or Artificial Neural Networks and bear the potential to deliver realistic and quantitative risk analyses relying on data rather than subjective interpretation or assumption. This may also hold for fuzzy data integration techniques.

However, most terrestrial modelling concepts have still difficulties to assimilate probabilistic statements about spatial heterogeneity of the available amount of observations or the occurrence of spatial and functional boundaries when expressed in the shape of likelihoods.

Hier fehlt ein ueberleitender Satz zum naechsten Kapitel.

Dieses Kapitel muss u.U. noch untergliedert werden. Es ist zu lang

FUTURE DEMANDS AND RESEARCH FIELDS

The recognition of patterns, either in the form of boundaries or structural units of high internal consistency, in spatial and temporal dimension, in the survey or model area, is one of the fundamental objectives when analysing Earth scientific data.

Algorithms that perform rapid and automated pattern recognition with no restriction to specific data types or any combination of data types are of high interest to the Earth science community. These algorithms should ideally offer a high degree of intuitive control to avoide highly abstract and complex hidden operations based on subjective user initialisation, for example, on a training data subset.. Only this kind of objective identification of structures in complex databases may help tailoring efficient and automated model parameterization for terrestrial process simulation. Identified pattern may also indicate different functionality of terrestrial processes between different structures assisting to set up suitable spatial or temporal compartments for process modelling (Kolditz et al. 2012).

, Algorithms should pay attention to the general characteristics of the available data pool when analysing and integrating disparate data . For example, a fundamental differentiation between subjectively sampled information, such as visual inspection of drilled material or soil classification, and sensor-controlled sampling data should be made. Subjectively sampled information may be perfectly right or wrong, depending on the experience and knowledge of the human analyst. Analysis of the subjectively sampled data does not allow for any objective and quantitative statement about the quality of the acquired information. In contrast, , sensor-controlled data are always limited in accuracy and information detail.

In many cases data analysis allows for quantitative or probabilistic statements about data quality, e.g. analysing noise levels and their spatial variations on a distinct scale of resolution. When integrating subjective and sensor-controlled information, e.g. mapped geology and airborne geophysical information, the mapped geology may be accurate in some regions and should guide the pattern recognition to overcome accuracy limits of the technical data, while in other regions the technical data should dominate the mapped geology where it may be imperfect. Developing and implementing algorithms that are capable of integrating data while paying attention to nature and quality of the observations is still a key challenge. These areurgently required to increase the acceptance of pattern recognition based data

integration approaches.

For convenience, pattern based data integration should provide the result of data integration and information extraction as a dimensionless numerical structural matrix expressing structural similarity or heterogeneity of the analysed database . Not only the most likely pattern should be provided, but also an objective assessment of the trustworthiness of the identified pattern is to be delivered. A simple technique is, for example, the description of the database heterogeneity by fuzzy membership information when using fuzzy cluster analysis for structural data integration (Figure 1) Since fuzzy cluster analyses do not pay attention to the nature of the considered input data, the fuzziness of the integrated information can only be regarded as a kind of internal classification consistency, but not as a quantitative probabilistic assessment of the identified pattern. Even Bayesian or Artificial Neural Network techniques can only come up with objective probabilistic statements about the integrated structural pattern when taking nature and uncertainty of the analysed data sets quantitatively into account. No techniques are currently known in Earth sciences that allow for this kind of complex and realistic structural integration. A few individual aspects, such as consideration of data noise or general provision of probabilistic quantification of detected features, are already available.

When integrating and trying to express the entire information of a geoscientific database in terms of a dimensionless structural information matrix under uncertainty, then the individual types of underlying information previously entered, e.g. natural gamma radiation intensity, mapped geology, or chemical element abundance information, can be considered as attribute information giving the dimensionless pattern description a geoscientific explanation . Information loss could be quantitatively judged by using the structural information as weighting scheme for the spatial reconstruction of the attribute information (Figure 1).

Going ahead, this could pave the way for reliable optimal sampling point identification when dealing with spatially continuous and sparse data sets. By first integrating the spatially

continuous information, the resultant dimensionless structural information could be analysed and rules could be inferred where to collect the sparse information. Fundamental assumption would be that at least some of the integrated spatially continuous data sets are somehow related to the sparse target parameter. In turn, the sparse data could be interpolated using abstract pattern descriptions resulting in interpolated maps with data-driven optimal spatial complexity or stochastic generations. The latter requires the integrated structural information to be formulated in terms of probabilities.

In practice, such approaches would require the design of hierarchical experiments, e.g., by acquiring and analysing the spatially continuous information prior to the collection of additional sparsely sampled data. Rapid data integration and analysis methods would be required to achieve acceptance of these hierarchical field setups.

Research funding - often aiming on periods between 2 and 5 years - further increases the need for rapid data integration techniques when aiming at hierarchical survey design. Thiswould implicitly mean for terrestrial modelling to cope with structural information and probabilistic statements when parameterising models and defining different spatial and temporal modelling compartments. Only this kind of approach would enable to make full use of quantitatively integrated data, thus ensuring an equal and realistic contribution of all available information to set up various model functionalities. The ncorporation of quantitatively integrated databases would enable the prediction of more realistic scenarios.

LINK TO OTHER RESEARCH FIELDS

In recent years, a number of different science and technology disciplines have experienced the phenomenon of ever-increasing acquisition amount and speed of information, e.g. security control, navigation, biology, economy, etc. For example, the last two decades saw a rapid development of (almost) real-time image analysis in security technology, largely computer-based genome de-sequencing, speech recognition, and customer-specific

advertising. Particularly those disciplines facing huge amounts of observations to be analysed in short time build nowadays routinely on automated information analysis rather than on human expert knowledge. Breaking the available information up into distinct pattern, e.g. for feature identification in image analysis or speech recognition, and interpreting the relevance of the identified pattern led to an algorithmic boom in data mining, machine learning and image processing techniques.

However, in a number of Earth science disciplines, a certain hesitation towards such automated information extraction of disparate data sets exists and human expert knowledge is preferred as it is judged unrivalledonsidering that Earth scientists who are interested in understanding the terrestrial environment even nowadays face huge volumes of data comprising physical, chemical, geological, ecological etc information, we believe that objective assessment of such databases by human interpreters is not possible. We believe that automated pattern based analysis tools providing integrated abstract quantification of informational heterogeneity of databases will support the rapid assessment of relevant information. Compared to other disciplines, Earth Sciences are definitely lagging behind in the utilization of these algorithms. The overwhelming majority of recent data integration methodologies introduced to Earth sciences going beyond the utilization of deterministic transfer functions, visual integration or classical geostatistical concepts have been inspired by algorithmic developments initially made for data types other than Earth scientific data. We see a high potential for increased information extraction from geoscientific databases when striving to benefit from algorithmic developments made in other scientific disciplines. However, for getting the most out of such new algorithmic developments, a high familiarity of Earth scientists with data mining, image processing and machine learning techniques is absolutely required. This kind of qualification will have to complement the geoscientific expertise. It will allow for better and objective information extraction when analysing large databases to understand and predict terrestrial environments under specific consideration of

uncertainties inherent to the available database.

ACKNOWLEDGEMENT

We thankfully acknowledge the support of the South African National Research Foundation (NRF) which provided the opportunity to run the two workshops as part of the South African-German Year of Science 2012-2013. We also express our gratitude to the audience present at both the workshops for insightful and interesting discussions and contributions which lastly initiated the draft of this paper.

REFERENCES

Kolditz O, Rink K, Shao H, Kalbacher T, Zacharias S, Dietrich P (2012) Data and modeling platforms in environmental Earth sciences. Environmental Earth Sciences 66:1279-1284

Kvamme KL (2006) Integrating multidimensional geophysical data. Archeological Prospecting 13:57-72

Wood EF, Roundy JK, Troy TJ, van Beek LPH, Bierkens MFP, Blyth E, de Roo A, Döll P, Ek M, Famiglietti J, Gochis D, van de Giesen N, Houser P, Jaffé PR, Kollet S, Lehner B, Lettenmaier DP, Peters-Lidard C, Sivapalan M, Sheffield J, Wade A, Whitehead P (2011) Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. Water Resources Research 47:W05301

Rink K, Kalbacher T, Kolditz O (2012) Visual data exploration for hydrological analysis. Environmental Earth Sciences 65:1395-1403

Zacharias S, Bogena H, Samaniego L, Mauder M, Fuß R, Pütz T, Frenzel M, Schwank M, Baessler C, Butterbach-Bahl K, Bens O, Borg E, Brauer A, Dietrich P, Hajnsek I, Helle G, Kiese R, Kunstmann H, Klotz S, Munch JC, Papen H, Priesack E, Schmid HP, Steinbrecher R, Rosenbaum U, Teutsch G, Vereecken H (2011) A network of terrestrial environmental observatories in Germany. Vadose Zone journal 10:955-973

FIGURE CAPTIONS

1: Toy example illustrating quantitative data integration based on pattern analysis. Two 2D spatial distributions (maps) of different data serve as input data base (Blue frame). The integrated information is described by a dimensionless abstract numerical 3D matrix. Here, this matrix is visualised as three 2D matrices. Additionally, attribute information is related to the structural information, e.g. in the form of mean values for each input parameter and class. The information in the black frames reflects the structural heterogeneity of both input models. Using the attribute matrix and the integrated structural matrix, both input models could be reconstructed (brown frames). Note the slight informational loss in amplitude, but the correct structural reconstruction.