Are Earth Sciences lagging behind in data integration methodologies?

Hendrik Paasche1,10, Detlef Eberle2, Sonali Das3, Antony Cooper3, Pravesh Debba3, Peter Dietrich1, Nontembeko Dudeni-Thlone3, Cornelia Gläßer4, Andrzej Kijko5, Andreas Knobloch6, Angela Lausch7, Uwe Meyer8, Ansie Smit5, Edgar Stettler9, Ulrike Werban1

1 UFZ Helmholtz Centre for Environmental Research, 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, Spatial Planning and Systems, P.O. Box 395, Pretoria 0001, South Africa

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

5 University of Pretoria Natural Hazard Centre, Africa, Private Bag X20, Hatfield 0028, Pretoria, South Africa

6 Beak Consultants GmbH, Am St. Niclas Schacht 13, 09599 Freiberg, Germany

7 UFZ Helmholtz Centre for Environmental Research, Department of Landscape Ecology, Permoserstr. 15, 04318 Leipzig, Germany

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

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

10 Corresponding author: telephone: +49 341 2351414; fax: +49 341 2351939
Email: hendrik.paasche@ufz.de

Hendrik Paasche • Detlef Eberle • Sonali Das • Antony Cooper • Pravesh Debba • Peter Dietrich • Nontembeko Dudeni-Thlone • Cornelia Gläßer • Andrzej Kijko • Andreas Knobloch • Angela Lausch • Uwe Meyer • Ansie Smit • Edgar Stettler • Ulrike Werban

This article reflects discussions German and South African earth scientists, statisticians and risk analysts had on occasion of two bilateral workshops on 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 events were arranged as part of the South African – German Year of Science 2012/2013. The South African National Research Foundation (NRF, UID 81579) has supported the two workshops as part of the South African – German Year of Science activities 2012/2013 established by the German Federal Ministry of Education and Research (BMBF) and the South African Department of Science and Technology (DST)..

Human welfare depends critically on sustainable utilization of our environmental resources. Earth Sciences are in part concerned with the observation of land-based terrestrial environments for the purpose of process understanding and of exploration of current states. 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 and management or hazard risk analysis. Observing and modelling terrestrial environments provides the knowledge base to handle a wide variety of societal issues such as economy regulation or migration.

To cope with the complexity of land-based 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 information 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. If attainable, data should be acquired using one sensor at different scales (OSADIS approach; Lausch et al. 2013). Such scale-bridging information about a distinct data type may help to overcome scale-related differences impeding comparative analyses and integration of disparate data.

When sampling information about our environment, Earth Sciences partly overlap with neighbouring disciplines. In some cases, Earth scientists may also be interested in assessing biologic diversity information, plant vitality, or intrusion/diminishing of species in a certain region to complement traditional geoscientific data. For example, plant vitality information may partly reflect variations in soil composition or chemical element concentration. The databases required to address the complexity of terrestrial environments may, therefore, 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, geochemical, or ecological context, respectively. Some distinct states of our environment may be satisfactorily addressed by a single sub-discipline. None of these individual disciplines will, however, provide all-embracing information about our terrestrial environment allowing for a holistic understanding of processes in the terrestrial environment, e.g. regional water balance development. When striving to assess the complexity of the terrestrial environment, Earth Scientists see themselves challenged by the task of analysing and integrating the information provided by various observation variables. This task ends up in the delivery of multi-method and multi-scale databases. Figure 1 summarizes the fundamental flow of information when modelling and predicting a number of processes or states of the terrestrial environment.

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 form of satellite imagery, geochemical and mineralogical sampling, geophysical survey data, ecological and hydrologic information, etc. The sheer size of the available amount of information 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, 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). The involvement of observed information into the modelling of processes going on in the terrestrial environment is critical, particularly, if “soft” data providing information that is non-linearly or even non-uniquely linked to the process of interest and the considered model parameters are considered. For example, when modelling subsurface water flow, geophysical information, such as electrical resistivity distribution, can be used to constrain the hydrological model. Moreover, in a deterministic sense, the exact relationship of the physical electrical resistivity parameter with hydrologic parameters required for hydrologic process modelling, e.g. with effective porosity, is usually not quantifiable. This relationship may also spatially vary because of changing sedimentary composition or evolutionary history of the deposited materials. Nevertheless, depending on the geoscientific discipline, a varying number of theoretical, empirical or semi-empirical deterministic transfer functions are frequently used to “convert” an observed data 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 number of sparse observations where the target parameter (porosity in our Archie’s law example) and the “soft” constrain parameter (spatial electrical resistivity distribution) are commonly known. Such a direct data integration approach relying on data conversion, particularly when going along with rather subjectively selected calibration sites, may impede high-quality modelling of terrestrial environments and even restrict the objective assessment of current states in these environments.

Some recent and flexible techniques of data integration have been implemented and introduced in the Earth Sciences. A number of critical issues need to be addressed by data integration techniques when applied to Earth Science databases. We are defining also a number of general future goals in this field of research that are considered of high and urgent relevance for the Earth Science community. The following outlines the demand for flexible, rapid and largely automated integrated analysis tools capable of keeping speed with an ever-increasing data acquisition rate. This requirement is not unique to Earth Sciences. In our view, consequences may arise from the rapid development of data mining and integration techniques to be considered in the training programs of young prospective Earth scientists. In the future, they will have to be in a position to analyse huge complex and interdisciplinary Earth scientific databases in a more objective and comprehensive manner than routinely done to the present day.

Data integration technologies in Earth Sciences

Earth Sciences have a long tradition in visual integration of disparate or complementary data sets (e.g. Kvamme, 2006; Rink et al. 2012). Traditional example is the use of a light table, which enables the semi-transparent overlay of several informational layers, i.e. maps. The modern approach is to use virtual environments to produce graphically integrated composites of different data sets. Popular examples are geographical information systems (GIS), which may be complemented by web-based services and multimedia contents suitable for realistic geovisualisation of dynamic environments (e.g. Gläßer et al. 2010). Another frequently used visual integration technique is the ternary plot. Co-located informational layers are integrated by scaling red, blue and green colour saturation according to the dynamic range of each informational layer. A fourth layer can be incorporated by scaling the brightness (grey tone) of the composite image.

Visual integration techniques are usually limited to a small number of informational layers, e.g. no more than four in the case of a ternary plot. The 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 trustworthiness extracted from the visually integrated composite images is not possible. Different interpreters may come up with analyses of the composites that are more or less concurrent. To some extent, the interpretational differences may reflect ambiguities inherent to the data acquisition accuracy. Imperfect technical understanding, subjective weighting of the importance of individual informational layers or misleading of the interpreters eye by different reception of blue, green and red colours in an image may cause an additional bias when integrating data visually. The utilization of visual data integration and information extraction prohibits consecutive objective risk analysis or reliable prediction of future scenarios. For instance, when subdividing a survey area into a number of internally consistent units (zones) based on visual data analysis, such subdivision may turn out to be more or less erroneous, as it depends on the skills of the human interpreter. The consecutive model parameterization building on the resultant zonation or definition of processes in the identified zones cannot be expected to be more appropriate than the initial spatial compartment identification based on subjective insights.

More powerful data integration strategies strive to go beyond simple visual integration by providing objective numerical information easing the integrated interpretation of multiple data sets. Geostatistical tools have come into use to analyse and describe spatial and/or spatiotemporal distributions of data. Concepts of random function theory are used considering the individual observations as well as missing observations (data gaps) close to the available samples as correlated random variables. The uncertainty with regard to spatial estimation and simulation of data can be described by a statistical model of spatial continuity, e.g. a variogram. Approaches exist, where statistical models generated for one data set are assumed valid for the prediction of other data sets. This is a more quantitative way of integrating data compared to visual integration and a more flexible approach compared to those utilizing an explicit deterministic transfer function, since uncertainty of data can be considered, e.g. in the form of different conditioned stochastic simulation realizations. Meeting the conditions allowing for a common statistical description, i.e. high correlation of different data sets, usually requires data processing and can only be performed to a certain limit of accuracy. 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 straightforward 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 patterns in visualised data. For example, one might be interested in identifying sub-areas exhibiting similar characteristics within one data set or between different ones. Conversely, one may be interested to assess the boundaries between such sub-areas. The identification of sub-areas may also pay attention to internal data characteristics, e.g., varying quality of observations reflected by different noise levels, the occurrence of distinct shapes in visualised data, e.g. lineaments reflecting dykes or faults, or circular features. 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 a spatially varying degree of sample correlation.

These techniques are in popular use in various degrees in the individual disciplines in Earth Sciences. Software modules integrating multiple informational layers and analysing the informational content in a probabilistic sense (using the above techniques) have been developed for specific tasks, e.g. the prediction of the mineral potential of a specific area (Fig. 2; Eberle et al. 2012; Noack et al. 2012). Data integration techniques that provide a probabilistic formulation of the results, e.g. based on Bayesian inference or Artificial Neural Networks, bear the potential to deliver realistic predictions in combination with quantitative risk analyses relying on *data* rather than subjective interpretation or assumptions. This may also largely hold for fuzzy data integration techniques (e.g. Hachmöller and Paasche 2013). However, most terrestrial modelling concepts have still shortcomings to assimilate probabilistic statements about spatial heterogeneity of the available observations or the occurrence of spatial and functional boundaries when expressed in the form of likelihoods or fuzziness.

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 without restriction to specific data types or any combination of data types are of high interest to the Earth Science community. Ideally, these algorithms should offer a high degree of intuitive control by avoiding highly abstract and complex hidden operations based on subjective user initialisation, e.g. by manual training or selection of a data subset for learning. 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, which assists setting up suitable spatial or temporal compartments for process modelling (Kolditz et al. 2012).

Algorithms should pay attention to the general characteristics of the available information 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 or quantitative statement about the quality of the acquired information. In contrast, sensor-controlled data are always limited in accuracy and informational 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 the accuracy limitations 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 capable of integrating data while paying attention to the nature and quality of the observations is still a key challenge urgently required to be solved in order to increase the acceptance of pattern-recognition-based data integration approaches.

For convenience, pattern-based data integration should provide the resultant integrated information in terms of a dimensionless numerical structural matrix expressing structural similarity or heterogeneity in the analysed database. Not only the most likely pattern should be provided, but also an objective assessment of the trustworthiness of the detected 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 (Fig. 3; Paasche and Eberle 2011). 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 recognised pattern. Even Bayesian or Artificial Neural Network techniques can only come up with objective probabilistic statements about the integrated structural pattern when taking the nature and uncertainty of the analysed data sets quantitatively into account. No techniques are currently known in Earth Sciences that allow for such complex and realistic structural integration. However, a few individual aspects, such as the consideration of data noise or general provision of probabilistic quantification of detected features, are already available.

When integrating and expressing the entire information of a geoscientific database in terms of a dimensionless structural information matrix under uncertainty, the individual types of underlying data sets, 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 meaning. Informational loss could be quantitatively judged by using the structural information as a weighting scheme for the spatial reconstruction of the attribute information (Fig. 3).

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 (e.g. Hachmöller and Paasche 2013). 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 (e.g. Hachmöller and Paasche 2013) 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 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. Current research funding practices – often aiming on periods between 2 and 5 years – further increase the need for rapid data integration techniques when planning to use a hierarchical survey design. For terrestrial modelling, this would implicitly require the need to cope with structural information and probabilistic statements when parameterising models and defining different spatial and temporal modelling compartments. 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. Such incorporation of quantitatively integrated databases would enable the prediction of more realistic scenarios when modelling terrestrial environments.

Link to other research fields

In the recent past, a number of different science and technology disciplines have experienced the phenomenon of an ever-increasing acquisition amount and speed of information, e.g. for security control, navigation, biology, economics 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 nowadays rely routinely on automated information analysis rather than on human expert knowledge. Classifying the available information into distinct patterns, e.g. for feature identification in image analysis or speech recognition, and for interpreting the relevance of an identified pattern has 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 appears to exist, and human expert knowledge is preferred as it is judged to be unrivalled.

Considering that Earth scientists who are studying the terrestrial environment currently face huge volumes of data comprising diverse 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 lagging behind in the utilization of such algorithms. The overwhelming majority of recent data integration methodologies introduced to Earth Sciences is going beyond the utilization of deterministic transfer functions, visual integration or classical geostatistical concepts. These methods have been inspired rather by algorithmic developments initially made for data types other than Earth scientific data. As progress is made in algorithmic developments in other scientific disciplines, we see a high potential for increased information extraction from geoscientific databases. However, for getting the most out of such new algorithmic developments, Earth scientists absolutely need to become thoroughly familiar with data mining, image processing and machine learning techniques. This kind of qualification is essential to complement the geoscientific expertise. It will allow for better and objective information extraction when analysing complex databases in order to understand and predict terrestrial environments under specific consideration of uncertainties inherent to the available database.

References

Eberle DG, Daudi EXF, Muiuane EA, Nyabeze P, Pontavida AM (2012) Crisp clustering of airborne geophysical data from the Alto Ligonha pegmatite field, northeastern Mozambique, to predict zones of increased rare earth element potential. Journal of African Earth Sciences 62:26–34

Gläßer C, Thürkow D, Dette C, Scheuer S (2010) The development of an integrated technical-methodological approach to visualize hydrological processes in an exemplary post-mining area in Central Germany. ISPRS Journal of Photogrammetry and Remote Sensing 65:275–281

Hachmöller B, Paasche H (2013) Integration of surface-based tomographic models for zonation and multimodel guided extrapolation of sparsely known petrophysical parameters. Geophysics 78:EN43-EN53

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

Lausch A, Pause M, Merbach I, Zacharias S, Doktor D, Volk M, Seppelt R (2013) A new multiscale approach for monitoring vegetation using remote sensing-based indicators in laboratory, field, and landscape. Environmental Monitoring and Assessment 185:1215–1235

Noack S, Barth A, Irkhin A, Bennewitz E, Schmidt F (2012) Spatial modeling of natural phenomena and events with Artificial Neural Networks and GIS. International Journal of Applied Geospatial Research 3:1–20

Paasche H, Eberle D (2011) Automated compilation of pseudo-lithology maps from geophysical data sets: A comparison of Gustafson-Kessel and fuzzy c-means cluster algorithms. Exploration Geophysics 42:275–285

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

Fig. 1 Fundamental workflow when striving to model and predict processes in the terrestrial environment. Note that only a holistic data analysis based on integrated consideration of multi-method and multi-scale data can offer reliable input information essential for realistic modelling. Ignoring information may result in biased input information for model setup.

Fig. 2 Schematic workflow for integrated data analysis and predictive modelling with artificial neural networks (ANN) and geographical information systems (GIS). Case study from 2010 by BEAK and Geological Survey Department of Ghana using different input geological units; distance to rock contacts, tectonic structures, intersection points, airborne geophysical data, and known training data (gold occurrences) are used for the prediction of gold mineralisation in NW-Ghana. The shown favourabilities on the predictive map range between 50 % (0.5) in green colour and 95 % (0.95) in red colour.

Fig. 3 Toy example illustrating quantitative data integration based on pattern analysis. Two 2D spatial distributions (maps) of different data serve as input database (*blue frame*). The integrated information is structurally 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.

Figure 1

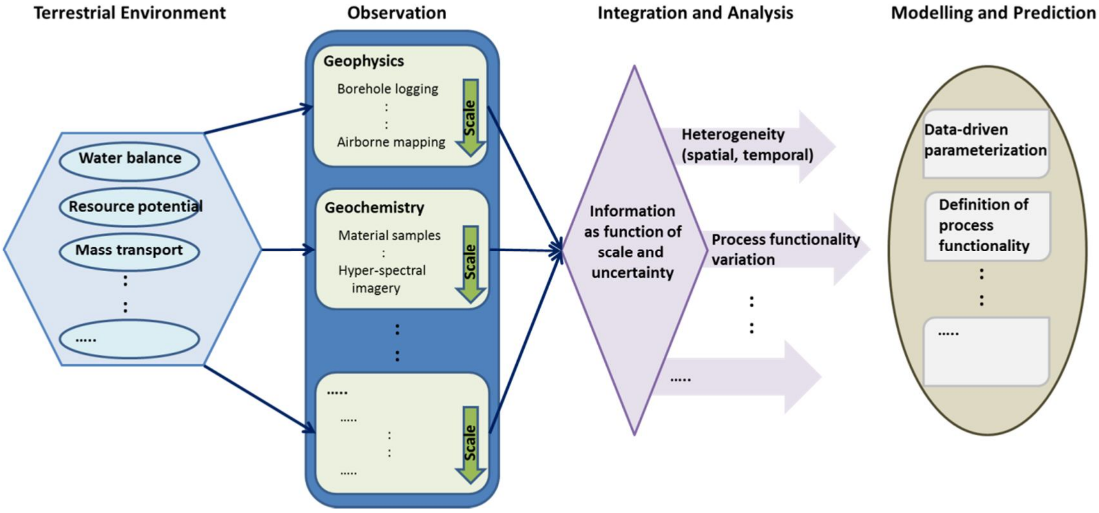


Figure 2

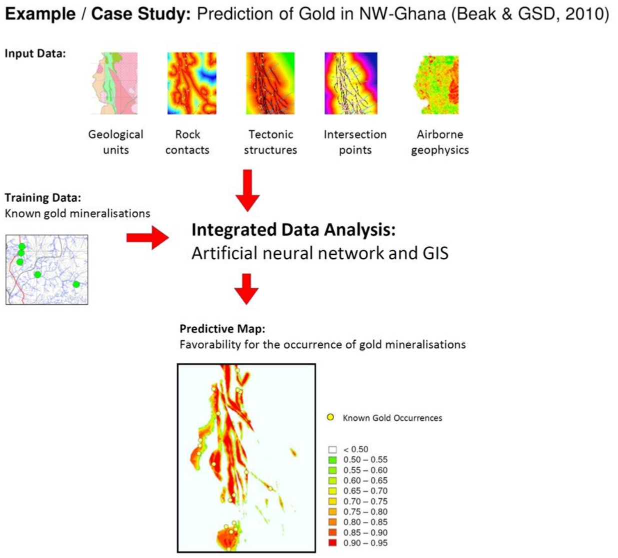
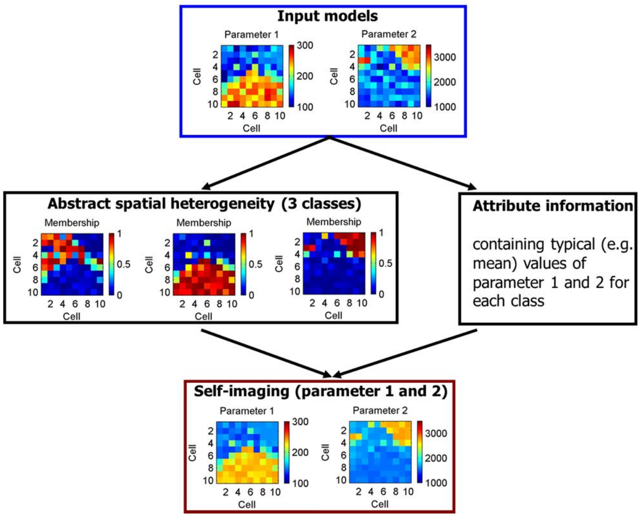


Figure 3



Input models
Parameter 1
Parameter 2
Cell
300
200
100
3000
2000
1000
Abstract spatial heterogeneity (3 classes)
Membership
1
0.5
0
Attribute information
containing typical (e.g. mean) values of parameter 1 and 2 for each class
Self-imaging (parameter 1 and 2)
Parameter 1
Parameter 2
2 4 6 8 10