VALIDATING MEASUREMENT DATA BY QUALITY MEASURES USING FUZZY-APPROACHES

Jochen Wittmann, Dietmar P.F. Möller
University Hamburg, Department for Informatics, AB TIS

Corresponding Author: J. Wittmann, University Hamburg, Department for Informatics, AB TIS
Vogt-Kölln-Straße 30, 22527 Hamburg, Germany, email: wittmann@informatik.uni-hamburg.de

Abstract. In many application situations, the researcher is confronted with the problem that there is a set of data sources available, e.g. measurements that refer to the same measured value, that substantially differ in data quality. Data quality in this context shall be defined along the criteria completeness (lost data, gaps, ...), scale (high – low), and precision (measurement error). Existing data explorations mostly concentrate on one of these criteria and offer proprietary solutions only. Those proprietary solutions are not capable to foster an integrative data interpretation that deals with measurements coming up from different data sources and measurement methods. However, putting together the data of different measurement approaches could bridge missing data and improve data reliability. The objective for this paper is to develop a data analysis and validation workflow that allows to combine the given strongly varying data sets to a common view on the resulting dimension.

To specify these tasks more precisely, the problem is formalized first giving every measured value an additional quality attribute. In consequence, the paper shows, how to use multiple measurements by different methods for improvement of the final result, how to incorporate the quality of the measurement into the mapping function, how to use fuzzy-methods for the classification task, and how to visualize the quality of the classification in the final 3D-representation.

Doing so, there are different levels of detail, the quality value can be specified: The most precise would be to attribute a distinguished value to every point (that means every tripel *(x,y,z)*) as implied by the formalism. For practical reasons, such an effort nor will be reasonable in respect to the time it would need to set a quality value for each point, neither it would be appropriate to the problem itself, because there is not that grade of differentiation in the measurements. Therefore, the level of differentiation will be determined by a common quality value for a certain measurement or even for the measurement method in general. This heuristic approach may be insufficient in respect to the granularity provided by the formalism and implies inaccuracies when different points are compared to each other within the data set of a single measurement, but for the overall interpretation of the data and in respect to the alternative interpretation pathways it is helpful to have a measure to compare the different data sets in regard on their contribution to the error minimization task in relation to the corresponding, competitive measurements. For practical reasons a general value for the measurement method will be pre-set that might be changed by the people who executed the measurements on level of measurement. For outliers the quality value can be overwritten explicitly on single point scale.

The paper provides in its first part the proper formalisation of the problem as scetched so far in this abstract, the second part will give a practical “to-do”-advice on the base of the MatLab Fuzzy-Toolbox and a self-developed collection of m-files and c-routines supporting the analysis and validating workflow to merge different measurements to an unique classification result.

1. The Problem Description

The problem is well known from many investigations based on environmental measurements [1], [5]: There are great differences in the quality of the data measured, for certain points in space there are no data available at all, in other segments of the 3D-space, however, data from different measurements is available and has to be assessed to gain a single result value for the corresponding points in space. In this situation, the general task splits into two subtasks: Firstly, the visualization task to elaborate the 3D-model (e.g. of the soil layers, of concentration layers in the air, or other structures found in the measurement data) from incomplete and uncertain data at all, and secondly the task how to integrate fuzzy-defined knowledge and data quality into the visualization task to illustrate the confidence level the method assigns to the values visualized.

2. Formalization of the Visualization Task

In these contexts as well as in many other situations, the researcher is confronted with two main problems: There is already a software solution that is highly specialized for evaluation and visualization data generated by the related measurement method and its corresponding hardware. However, the data format is proprietary and the methods offered are restricted [7]. Those proprietary solutions are not capable to foster an integrative data interpretation that deals with measurements coming up from different data sources and measurement methods. However, putting together the data of different measurement approaches could bridge missing data and improve data reliability.

To solve this integration and interpretation problem, a mathematical formalization is helpful, because it lifts the problem on an abstract level that shows the structural interdependencies. This is done here in the context of the HADU project [1], but it might be generalized easily and could be adapted to any other visualization problem of environmental data.

To find structures in the data, for every point in space a correspondent mapping or classification to a set of predefined classes is necessary:

$$\begin{array}{llll} (x, y, z) \rightarrow i & \text{with} & x, y, z\ \varepsilon\ R & \textit{space coordinates} \\ & & i\ \varepsilon\ N & \textit{class} \end{array}$$

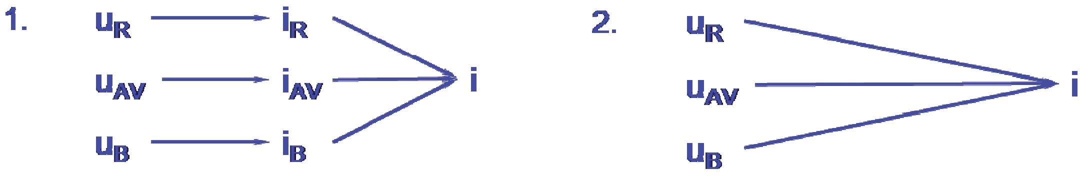
The measurements can be represented by functions of the following type, with X for the different data sources:

$$\begin{array}{ll} F_X :\ (x, y, z) \rightarrow u_X \quad \text{with} & x\ \varepsilon\ D^X_x \subseteq R \\ & y\ \varepsilon\ D^X_y \subseteq R \\ & z\ \varepsilon\ D^X_z \subseteq R \\ & u_x\ \varepsilon\ R \end{array}$$

For the beginning, this formulation for the measuring functions neglects the fact that the measurement datasets will be incomplete in the x,y,z-domain and that there might be only few categories for the value of u instead of the continuous range of *R.*

3. Specification on Base of the Formalism

In this situation, we have a measured value for each position but for visualization and interpretation a mapping between the measured value and the classes observable is necessary. To get the intended classification for visualization, there are two basic alternatives:



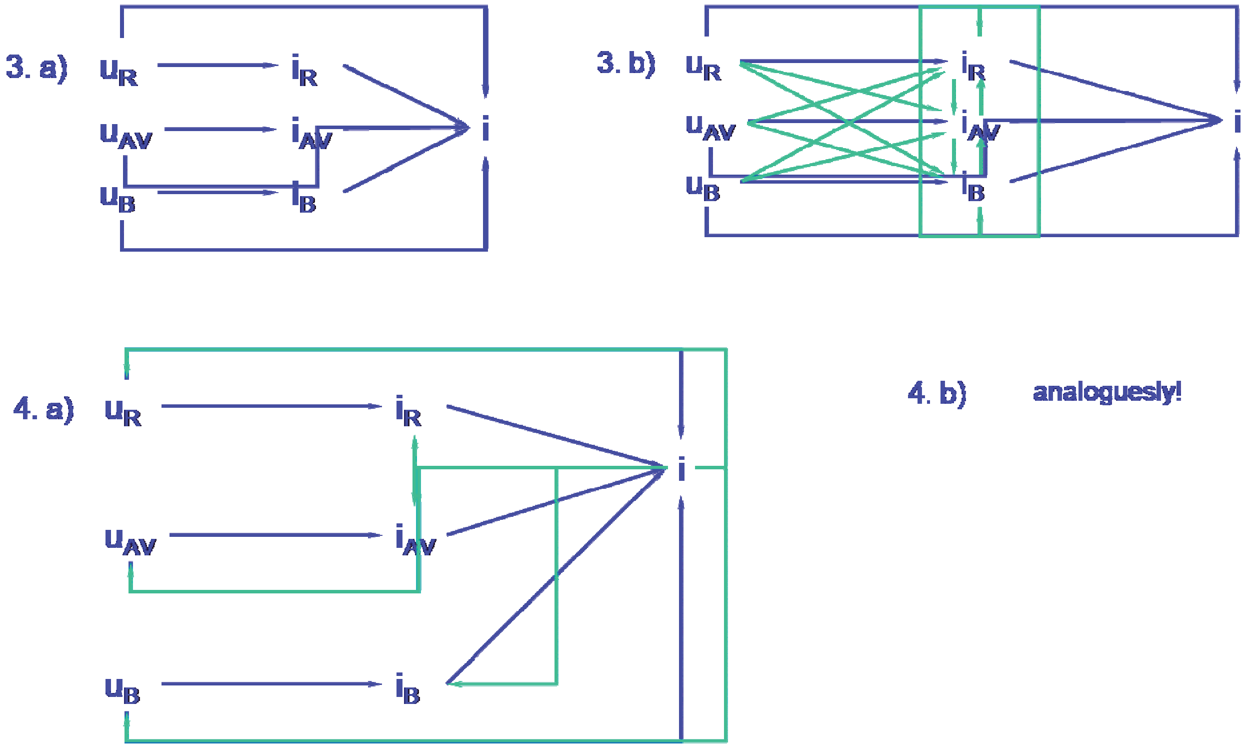


Fig. 1: unification/interpretation variants

1. Do the mapping for each measurement function *u* first, and combine the results from this step later to a common interpretation. (fig. 1.1)
2. Combine the different measurements directly to the final classification. (fig. 1.2)

These basic versions can be modified by recursive and interactive mapping variants:

3. Mixed classification: Individual classification for each measurement method first, integration to an over-all classification by integrating raw data and individual classification results later on. (fig. 1.3a) Or: Integration of all raw data sets for classification of the individual measurement results and the final integrative classification i. (fig. 1.3b)
4. Feedback classification: Execute classification as described in alternative 1. to 3., but try to adapt the individual classifications to the overall classification by a feedback calculation. Thereby, the depth of the feedback may be varied: depth 1 means feedback to the level of individual classifications, depth 2 means feedback to the level of measurements. (fig. 1.4a) This procedure can be modified by integrating the data of other measurement methods for individual classification as in case 3b, what leads to a complete feedback network between data and interpretations in analogy to fig. 1.3b.

A summarizing evaluation of these four variants leads to the following characteristics: The straight forward approach seems to be easy, because it requires only knowledge on the workflow of the measurement method currently under consideration. On the other hand, there is no support for the mapping task by additional data coming from additional measurements. The integrative analysis in the mixed and feedback alternatives allows benefiting from other data sources but it demands for rules how to integrate these data and especially how to weight contradictionary interpretations.

In practice, there are very heuristical approaches mostly based on mathematical methods available instead of a goal driven workflow for interpretation. In contrary to this, this paper emphasizes the formalism to force the users to give a precise semantical meaning of every step in the interpretation workflow.

In general, the intention of all these variants is to close gaps in the measured data and to achieve better mapping quality. Therefore, the mapping function has to solve three problems that can be handled by the listed more or less known methods (fig.2 outlines these major tasks):

- incomplete data:	interpolation
- redundant data:	error minimization
- uncertain data:	fuzzy classification and fuzzy interpolation

With these formalisms, the initial tasks can be formulated more precisely in the form of positive callings to give constructive answers or at least hints for solution:

1. Use multiple measurements by different methods for improvement of the final result!
2. Incorporate the quality of the measurement into the mapping function!
3. Visualize the quality of the classification in the final 3D-representation!
4. Use fuzzy-methods for the classification task!

Doing so, there are two more inputs necessary, the user has to provide: quality information on each measurement and the specification of the fuzzy classificator by member function, rules and defuzzyfication function. [6]

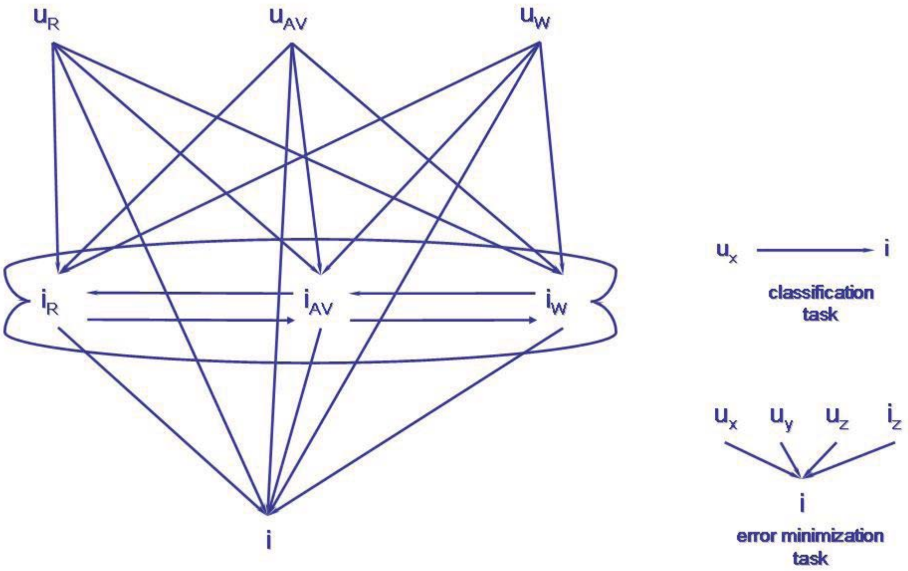


Fig. 2: basic tasks to generate a generalized interpretation

4. Integrating a Measure for Data Quality

So far, the way from raw data to the interpretation has been described. To integrate data quality, two dimensions have to be observed: On the one hand the quality provided by the raw data itself, on the other hand the quality or better to say the error caused by the different functions and methods on the way from the raw data to the final classification and interpretation. With these deliberations in mind, the quality can be represented within the introduced formalism by an additional value for the function f and the resulting classification function k as follows:

$$\begin{array}{llll} f^Q \;:\; (x, y, z) \rightarrow (u, q) & \textit{with} & x,y,z,u & \textit{analogue } F_X \\ & & q \;\varepsilon\; Qual & \textit{quality range} \end{array}$$

$$\begin{array}{llll} k^Q \;:\; (u, q) \;\rightarrow\; (i, q) & \textit{with} & u, i & \textit{analogue } F_X \\ & & q \;\varepsilon\; Qual & \textit{quality range} \end{array}$$

With this definitions, we can look back to section 3 with the different alternatives for the data evaluation workflow and are able to set up a more algorithmically formalization for the alternatives. All the variants that use feedback or balancing pathways have to be realized by a loop structure in the form of iterative recalculation of the classification and the error minimization algorithm until an acceptable quality level will be reached.

To complete the discussion, the definition of q as a value of the quality range Q has to be rendered more precisely. There are different possibilities to realize the quality attribute q:

- qualitive value
- fuzzy value
- epsilon interval for the given value
- quantitative value

And there are different levels of detail, the quality value can be specified: The most precise would be to attribute a distinguished value to every point (that means every tripel (x,y,z)) as implied by the formalism. For practical reasons, such an effort nor will be reasonable in respect to the time it would need to set a quality value for each point , neither it would be appropriate to the problem itself, because there is not that grade of differentiation in the measurements. Therefore, the level of differentiation will be determined by a common quality value for a certain measurement or even for the measurement method in general. This heuristic approach may be insufficient in respect to the granularity provided by the formalism and implies inaccuracies when different points are compared to each other within the data set of a single measurement, but for the overall interpretation of the data and in respect to the alternative interpretation pathways it is helpful to have a measure to compare the different data sets in regard on their contribution to the error minimization task in relation to the corresponding, competitive measurements. For the HADU system a general value for the measurement method will be pre-set that might be changed by the people who executed the measurements on level of measurement. For outliers the quality value can be overwritten explicitly on single point scale.

5. Resulting Workflow and Design Decisions

For the final user the workflow for integrating the different measurements and assessing them, is crucial for the final data quality. There are different sequences and interdependencies possible, which influence the intended final 3D-visualization. Figure 3 depicts the situation for the example of the HADU-project and will be interpreted according to the different steps in the workflow.

The figure starts on the left side with the situation in the domain measured (domain 3D), continues with the raw data of the measurements (radar R, ambient vibrations AV, and wells W). These measurements have to be interpreted to achieve the classification info, a step, which leads to the partial interpretation results IR, IAV, and IW or to an interpretation I. Analogously, there are different ways to visualize the interpretations: independently on each other, as integrative 3D-model calculated on base of the partial interpretations, or without any interpretation based on pure raw data.

For the HADU project, the different pathways are still in discussion. Two main lines can be recognized in the current state of the work: From the point of view of the users the visualization is intended as soon as possible to get a “feeling” of the data measured. The commitment for an interpretation is postponed as far as possible to the end of the process. In contrary to these user demands, it seems optimal for good over-all visualizations to integrate data on the base of existing and validated partial interpretations.

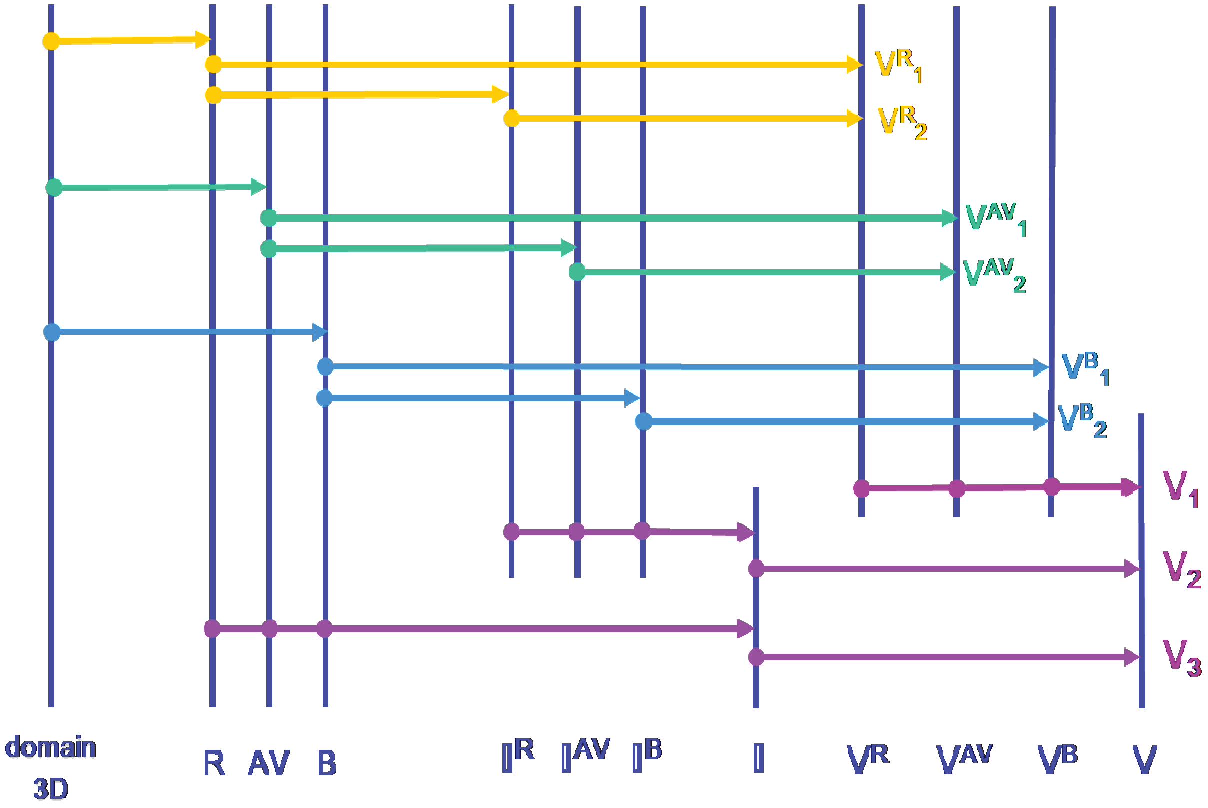


Fig. 3: workflow from raw data to interpretation and visualization

6. Treatment of Uncertainties by Fuzzy-Approaches

The preceeding sections introduce a special attribute that describes the quality of the measurement, i.e. a valuation for each point in space how accurate the measured value is estimated. So far, this attribute is introduced as a real value attribute for each of the points in space under observation. For practical reasons, however, it is difficult to determine the exact quality of a measurement integrating the different measurement devices and the context of the measurement campaign. An exact quality value for each point will not be practicable, therefore. In contrary to the preceding theoretical section, the following example shall explain the fuzzy-set-based approach to realize a qualitative evaluation by using the functionalities of the fuzzy-toolbox of MatLab.

The workflow for the user is supported by the toolbox in five steps as follows ([3], [4]):

- FIS Editor
- Membership Function Editor
- Rule Editor
- Rule Viewer
- Surface Viewer

The first step is to specify the dimension of the problem and the corresponding fuzzy system. In the example, there are three sets of measurements each with an additional and separate quality set, i.e. 6 inputs as depicted in figure 4.

In the second step the membership functions have to be specified. For each type of measurement the classification according to the interpretation function i is demanded. Additionally a classification for the quality of the measurement is necessary: Here, the workflow in practice differs from the theoretical one by assuming only a set of possible quality values such as “good”, “medium”, and “bad”.

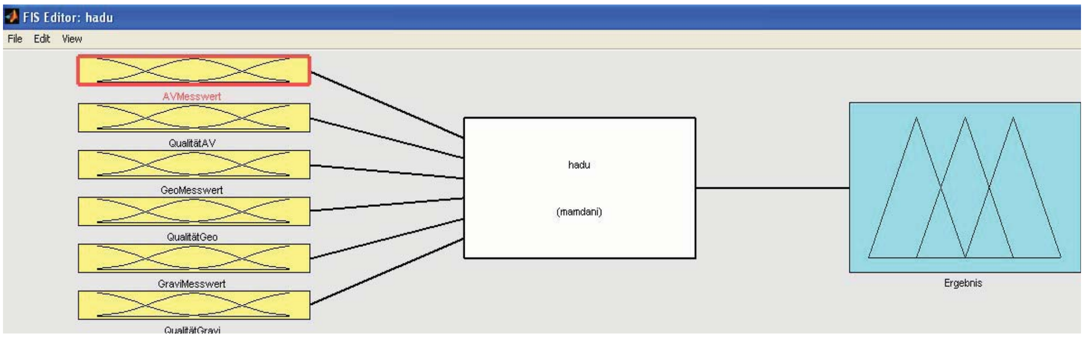
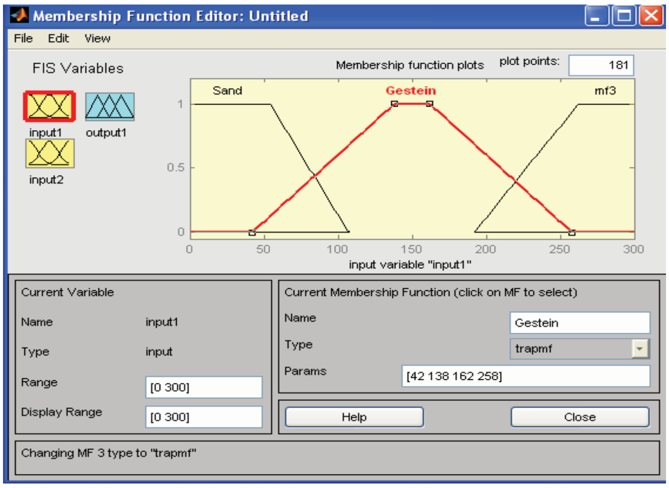


Fig. 4: fuzzy evaluation system



plot points: 181

Current Variable

Name: input1

Type: input

Range: [0 300]

Display Range: [0 300]

Current Membership Function (click on MF to select)

Name: Gestein

Type: trapmf

Params: [42 138 162 258]

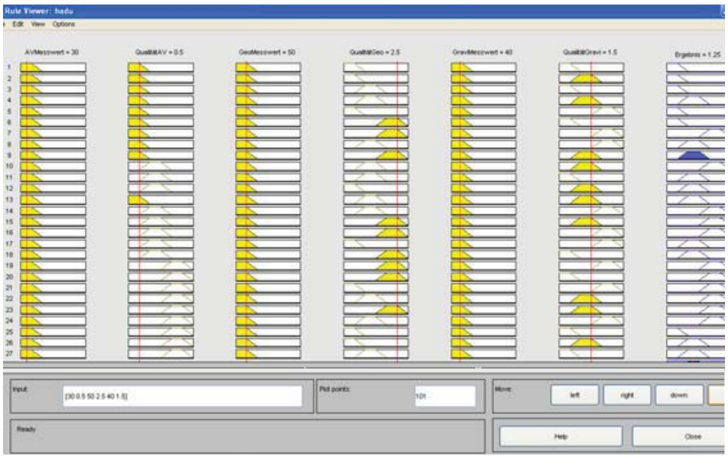
Changing MF 3 type to "trapmf"

Fig. 5: specification of the classification

The most important step is defining the rules for final classification within the rule editor: In this step, the user has the choice to combine the interpretation of the measurement values with the corresponding quality values. As explained in the theoretical section of the paper, such a valuation can be applied for the interpolation phase as well as for classification phase. The theoretical concept and the toolbox support both ways, a general decision and suggestion in this point can not be made because it depends on the individual objectives of the analyze and the data available.

Independent of the fact whether the rule set is designed for interpolation or for classification purposes, it has to be emphasized, that in both cases, not only an interpolation or a classification result has to be derived. In addition, for every value treated a corresponding quality value has to be provided from the fuzzy system and therefore a corresponding rule set to derive these quality measures must be developed in correspondence to the value generating rules.

Especially to explore the influence of these rules, a rule viewer is offered by the MatLab toolbox (see fig. 6). This tool marks the rules that are active when a certain input is offered to the system. Doing so, the rule viewer supports the users to develop a complete and consistent set of rules, but it does not show the effects of the rules on the final objective, the interpolation or the classification. To give the users an additional instrument to get a “picture” of the problem and the resulting classification, a special visualisation tool has been developed that is sketched in the following section.



Input: [30 0.5 50 2.5 40 1.5]

Plot points: 101

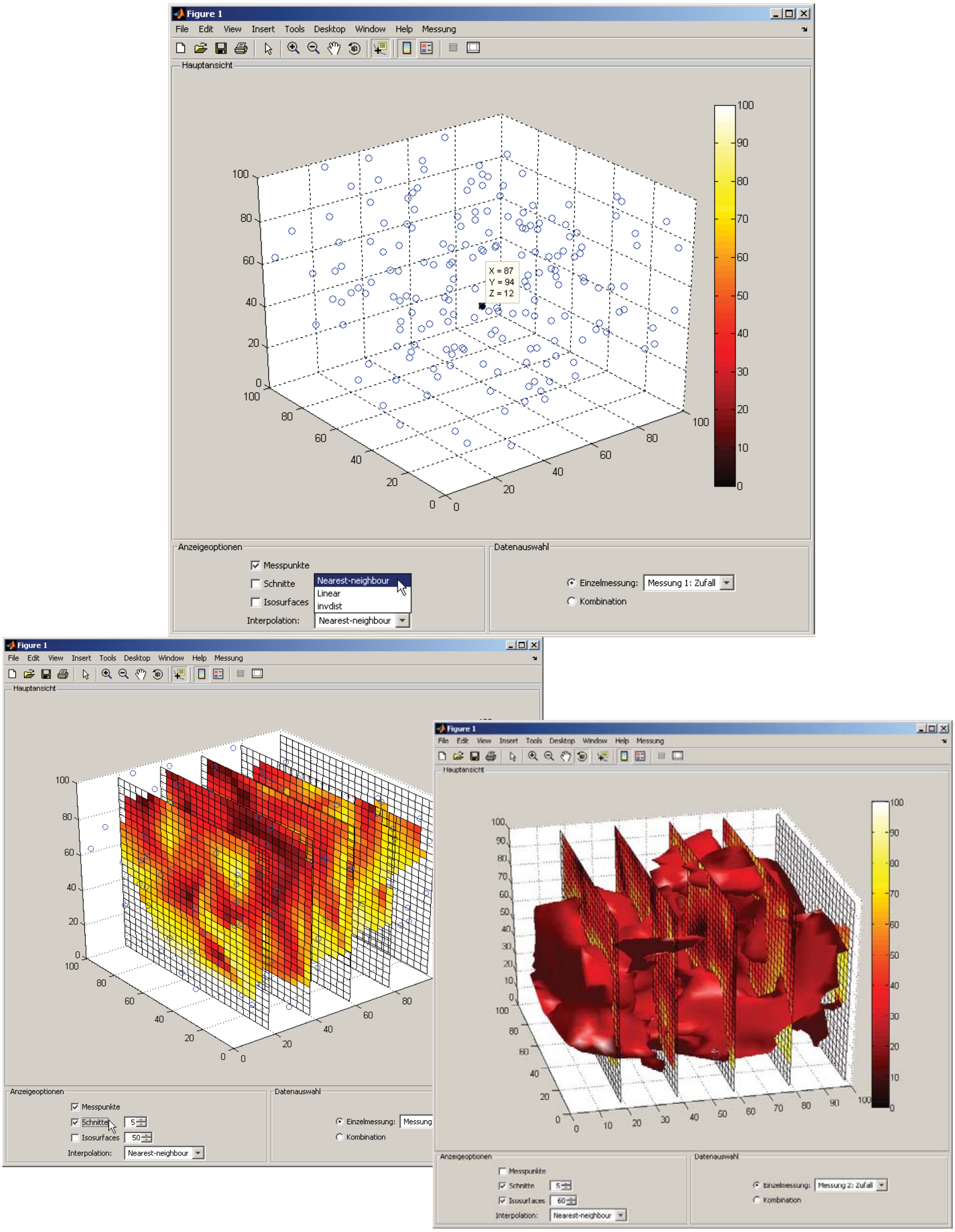
Ready

Fig. 6: The rule viewer

7. The Visualization Tool

The visualisation tool offers beside of the fuzzy component the regular methods for interpolation such as nearest neighbour, linear, ... to make comparisons between the fuzzy-based approach and the standard methods possible. However, main intention for the visualization tool is to give the user an interactive and stepwise support to get an impression of the data set under consideration and the model derived from these data. The visualization component is able to work on the data format of the fuzzy system in every stage of the workflow.

Figure 7 shows the interface. On the right side the selection of the data is offered, that are visualized separately or in a common view. Thus, integration of different measurements in a single view (and a single model) is possible. The measurements are marked by circles, the can be evaluated in the sense of interpolation or classification by one of the methods offered in the pull-down menu on the left side. To explore the data, iso-surfaces and intersection planes can be defined interactively. In complete analogy to the data evaluation process, a second window contains the identical view for the quality attribute. If both windows are used, changes in the viewpoint for the 3D-model are synchronized, giving always an identical perspective on measurement and quality data.



☑ Messpunkte
□ Schnitte
□ Isosurfaces
Interpolation: Nearest-neighbour

● Einzelmessung: Messung 1: Zufall
○ Kombination

☑ Messpunkte
☑ Schnitte 5
□ Isosurfaces 50
Interpolation: Nearest-neighbour

● Einzelmessung: Messung
○ Kombination

□ Messpunkte
☑ Schnitte 5
☑ Isosurfaces 60
Interpolation: Nearest-neighbour

● Einzelmessung: Messung 2: Zufall
○ Kombination

Fig. 7: The visualization tool: interface, intersection planes and iso-surfaces

8. Conclusions

Thus, the paper emphasizes the value of a proper mathematical formalisation as base for discussion and interpretation of environmental data that normally are incomplete and uncertain. The formalisation not only allows a compact description of the problem, but also facilitates a correct mathematical and algorithmic treatment and an efficient approach to calculate the high dimensional and large-array problem with a tolerable amount of CPU-time.

It offers:

a) A framework to build own interpretation and visualization workflows to combine different data sets.

b) A base for discussion of the different methods for classification and error minimization in well defined and clear-cut steps with emphasis on their interdependencies.

c) The opportunity to integrate knowledge on the system under consideration step by step and in scalable manner (as exemplarily explained for the granularity of the quality value).

However, it shows that the apparently simple task of data visualisation implies serious problems concerning data interpretation that cannot be solved automatically even not by the most elaborated and sophisticated visualization tools. All the steps between raw data and ostensive visualization have to be masterminded by lots of knowledge, which has to be elaborated from the experts sometime.

9. Links and References

[1] Geotechnologien, Science Report No. 6, BMBF 2005, S.124-140

[2] Gnauck, A., Luther, B. (2005) *Zur Auffüllung von Datenlücken in Zeitreihen der Wassergüte*, in Wittmann, J., Thinh, N.X. Simulation in Umwelt- und Geowissenschaften, Workshop Dresden 2005, Aachen 2005, pp. 295-305

[3] http://www.mathworks.com/products/matlab/

[4] http://www.mathworks.com/products/fuzzylogic/

[5] Reuther, C.D., et al. (2005) *Hamburg – A Dynamic Underground (HADU) Subsurface Evaluation of Hamburg Based on Analysis and Modeling of Recent Geological Structures and Dynamic Processes* in Geotechnologien, Science Report No. 6, BMBF 2005, pp.124-140

[6] Wieland, R. et al. (2004) *SAMT - eine neue Open Source Plattform zur Lnadschaftsanalyse, Modellentwicklung und Integration räich expliziter ökologischer und ökonomischer Modelle* in Schiefer, G. et al.: Integration und Datensicherheit - Anforderungen, Konflikte und Perspektiven, Referate der 25. GIL Jahrestagung, 8.-10. September 2004, Bonn, 2004, pp. 137-140

[7] Wittmann, J. (2007) *A Software Architecture for the Cooperation Project HADU: Hamburgs Dynamical Geological Underground* in Proceedings of the EnviroInfo 2007, Warsaw