Raport Badawczy
Research Report

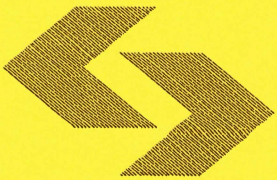
RB/9/2014

Statistical methodology for verification of GHG inventory maps

J. Verstraete, Z. Nahorski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. dr hab. inż. Zbigniew Nahorski

Warszawa 2014

Appendix 1: Solving the map overlay problem with a fuzzy approach, supplement

Verstraete Jörg[1]

[1] *Systems Research Institute - Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Polska.*

(email: jorg.verstraete@ibspan.waw.pl, tel: +48-22-3810100)

Abstract

The map overlay problem occurs when mismatched gridded data needs to be combined; the problem consists of determining which portion of grid cells in one grid relate to partly overlapping cells of the target grid. This problem contains inherent uncertainty, but is an important and necessary first step in analysing and combining data; any improvement in achieving a more accurate relation between the grids will positively impact the subsequent analysis and conclusions. Here, a novel approach using techniques from fuzzy control and artificial intelligence is presented to provide a new methodology. The method uses a fuzzy inference system to decide how data represented in one grid can be distributed over another grid using additionally available knowledge; thus mimicking the higher reasoning a human would use to consider the problem.

This supplement contains the technical description on how the inference system used in the main article was developed.

S1. Defining the inference system

S1.1 Concept

To refit data supplied on one grid to another grid, an inference system will be used. The approach is not to make any assumptions on the data itself, but rather to simulate a reasoning on the data as was illustrated in Section 2.3 of the article. In order to develop the inference system, a simpler example as shown on Figure S1 in this supplement will be used. There are two grids: one with two grid cells, and one with three grid cells. This is a very simple example where the grids cover exactly the same region of interest, and where the output grid and the grid containing extra information are the same, but it allows us to reason about the problem. The question of redistributing grid A translates to: which portion of B_2 contributes to A_1 and which portion to A_2? In other words, how do we split B_2 in B_{2a} and B_{2b} so that everything makes sense?

S1.2 Defining the parameters

The first step in defining the intelligent system is defining the parameters and the rules. First, it is necessary to determine which cells play a part in defining the value of an output cell. To do this, the relations between the values in the cells will be derived. Consider that all the values in all grid cells are constant, and that only the value in A_1 increases from 100 to 200. The effect of this increase is that the proportions between the values of the input grid A and the auxiliary grid C change from $\frac{100+100}{100+100+100}$ to $\frac{200+100}{100+100+100}$. This implies that the value

$f(C_1)+f(C_2)$ should increase from 100 to 200, in order to keep a correct proportion. In the sub cells C_{2a} and C_{2b}, it can be seen that the value $f(C_{2b})$ is reduced to 0, whereas the value of $f(C_{2a})$ has to increase to compensate for this. As a result, it can be concluded that a change of $f(A_1)$ has a proportional effect on $f(C_1)$ and $f(C_{2a})$ and an inverse proportional effect on $f(C_{2b})$.

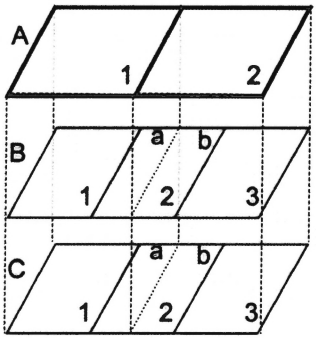


Figure S1. Simplified example using two grids: the output grid *B* and the auxiliary grid *C* use the same raster. The cover the same region of interest the input grid *A*, but the gridcells do not overlap nicely: grid *A* has two gridcells, whereas grids *B* and *C* have three gridcells.

Using similar reasoning on more complex cases, it was found that the cells that exhibit a proportional relation to a given output cell are:

- the cells of the input grid that intersect with the output cell
- the cells of the related grid that intersect with the output cell

The cells that show an inverse proportional relation to a given output cell are:

- the cells of the auxiliary grid that do not intersect with the outputcell, but do intersect with the inputcells that intersect with the output cell
- the cells of the input grid that do not intersect with the outputcell, but do intersect with the cells of the auxiliary grid that intersect with the output cell

In order to determine the value of the new cell, these relations will be used. As parameters, proportional and inverse proportional data from both input grid and auxiliary grid will be considered, resulting in 4 parameters for the rulebase. Unlike in the above example, the rules will not reflect any quantitative connection between the input grid and auxiliary grid: the proportion of the values in input grid and auxiliary grid is not taken into account when calculating the values for the output grid. Using that knowledge can allow a further refinement of the method, but to verify if the proposed method works at all, the current approach is sufficient.

Appendix 2 concerns different possible parameters; automatically determining which parameters are good for a given input is covered in Appendix 3.

S1.3 Defining the rules

Consider some possible cases as listed in Table 1.

Table S1. Example values for the conceptual example to reason about the gridcells.

	case 1		case 2		case 3		case 4	
cell	value	label	value	label	value	label	value	label
A_1	100	high	100	high	100	high	100	medium
A_2	100	high	100	high	0	low	200	high
B_1	100	high	0	low	100	high	0	low
B_2	100	high	100	high	100	high	150	high
B_3	100	high	100	high	0	low	0	low
B_{2a}	50	medium	100	high	0	low	50	medium
B_{2b}	50	medium	0	low	100	high	100	high

- In case 1, all the values are 100. The associated values with the two emission cells are equal, and the associated values with the three auxiliary cells are equal. To split B_2, it therefore makes sense to have an even distribution: B_{2a}=B_{2b}=50 (it implies that a total value of 300 in the auxiliary grid relates to a total value of 200 in the emissions, when considering the total value of both grids).
- In case 2, A_1=100 and B_1=0. This basically means that B_2 is the only grid cell contributing to A_1. At the same time, A_2=100 and B_3=100. If the total region of interest is considered, then there is a total value of 200 for all A_i together, and a value of 200 for all the B_j combined. Over the region of interest, this means that a value of 200 in the covariates relates to a value of 200 in the emissions; this is possible to maintain if B_2 is split such that it contributes completely to A_1, thus B_{2a}=100 and B_{2b}=0 (this could for instance be a situation where there is a point source in B_{2a}).
- Case 3 shows a situation where A_1=100 and A_2=0, and where B_1=B_2=100 and B_3=0. This basically implies that B_1 and B_2 are contributing only to A_1, hence B_{2a}=100. In this situation, a total value of 200 in the auxiliary grid relates to a total value of 100 in the emissions.
- The last case, case 4, shows A_1=100, A_2=200 and B_1=B_3=0; implying that B_2 contributes to both C_1 and C_2, but twice as much to A_2. Thus B_{2a}=50 and B_{2b}=100.

The above cases only use extreme values to illustrate the point, but similar reasoning can be done for intermediate values. Without resorting to calculations, similar results can be obtained by merely labelling the values with linguistic terms, case 4 then translates to:

> if A_1 is *medium* and A_2 is *high* and B_1, B_3 are *low* and B_2 is ***high***
> then B_{2a} is ***medium*** and B_{2b} is ***high.***

The linguistic terms can be modelled by fuzzy sets, which will be explained in section 1.4. Note that this method of reasoning does not depend on the shape of the grid cells. It also makes no assumption on how the data within a grid cell or across grid cells could be distributed. The areal weighting approach would not make any distinction between these cases, and always split C_2 evenly. It is obvious from the examples that this can be quite different from the actual situation. The example is too small to really show the areal smoothing method, but one can see that it is possible that cases 2 and 3 can be incorrectly assessed as there is no smooth transition; rather the contrary: in case 2, C_{2a} should be high because B_1 is low.

In order to derive the rules for this simple example, we first consider a number of extreme cases as shown on table 1. For ease of interpretation; all the values (both for auxiliary grid and input grid) are in the range 0-100. The first five rows show the known data

(data that will help determine the premises); the rows B_{2a} and B_{2b} show how B_2 should be distributed based on the known data (the conclusions of the rules).

To derive the rules, let us first assume that $B_1=B_3$ (as in cases 1 and 4). If $A_1=A_2$, then it is obvious that B_2 should be evenly split over both B_{2a} and B_{2b} (case 1). If $A_1<A_2$, it implies that B_2 contributes more to A_2 than to A_1; as a result $B_{2b}>B_{2a}$ (case 4). To make a rule that represents this case, we need to define the rule as:

$$\text{IF } A_1 < A_2 \text{ AND } B_1 = B_3 \text{ THEN } B_{2a} < B_{2b}$$

The output value clearly depends on the difference between A_1 and A_2: the greater this difference is, the smaller the value of B_{2a} should be. An analogue reasoning holds when $A_1>A_2$.

Next, assume the inputs are equal: $A_1=A_2$. If $B_1<B_3$, then it implies that, as emissions are equal, B_2 contributes more to A_1 than to A_2; so $B_{2a}>B_{2b}$; the greater the difference between B_1 and B_3, the more this should be reflected in the output. Consequently, this results in a rule:

$$\text{IF } A_1 = A_2 \text{ AND } B_1 < B_3 \text{ THEN } B_{2a} > B_{2b}$$

Again the greater the difference between B_1 and B_2; the more B_{2a} should differ from B_{2b}. A similar reasoning holds when $B_1>B_3$. In general, no values will be equal. This implies that rules for those mixed cases must be defined as well. In the rulebase however, values are not compared against each other but against predefined fuzzy sets. As a result, we first need to define what will be considered high and low. To define the rules, three predefined fuzzy sets for the grid A (representations for *low*, *normal* and *high: lowForInput*, *medForInput* and *highForInput*), three possible values for the auxiliary values (*lowForRelated*, *medForRelated* and *highForRelated*) and nine possible values for the outputted percentage (*outputLow4*, *outputLow3*, *outputLow2*, *outputLow1*, *outputNeutral*, *outputHigh1*, *outputHigh2*, *outputHigh3*, *outputHigh4*); all the fuzzy sets are explained in Section S1.3.4. The reason for the large number of output fuzzy sets is mainly that this number occurs natural when considering all possible cases and thus it allows for an easier generation of the rules. For all combinations of input values, an appropriate output value needs to be determined. Given that there are three possibilities for each input value, there are $3^4=81$ combinations and thus 81 rules. The rules in the rulebase are therefore of this form:

```
RULE 1 :IF inputProportional IS lowForInput
         and inputInverseProportional IS lowForInput
         and relatedProportional IS lowForRelated
         and relatedInverseProportional IS lowForRelated
         THEN output IS outputNeutral WITH 1 ;
RULE 2 :IF inputProportional IS lowForInput
         and inputInverseProportional IS lowForInput
         and relatedProportional IS lowForRelated
         and relatedInverseProportional IS medForRelated
         THEN output IS outputLow1 WITH 0.9 ;
RULE 3 :IF inputProportional IS lowForInput
         and inputInverseProportional IS lowForInput
         and relatedProportional IS lowForRelated
         and relatedInverseProportional IS highForRelated
         THEN output IS outputLow2 WITH 1 ;
```

In the system, each rule can be assigned a weight (indicated by WITH). The weights were chosen so that the rules that apparently contradict, i.e. those where input grids and auxiliary grid exhibit a different behaviour that contradicts the knowledge that both are

related, are assigned lower weights. From the experiments performed, this had almost no impact on the results.

S1.4 Defining the fuzzy sets

In the above sections, the linguistic terms *lowForIput*, *highForInput*, etc. were used. There of course needs to be a representation of these linguistic terms. The current prototype is designed to dynamically make definitions for these terms using fuzzy sets. For each output cell under consideration, the fuzzy sets for the linguistic terms that relate to the input are defined by considering the values of the input cells that play part in determining the value of this output cell. The values of the cells that exhibit a proportional relationship are summed; the values of the cells that exhibit an inverse proportional relationship are summed, and an interval between the minimum and the maximum of both values is defined. The fuzzy sets for low, medium and high are then defined on this interval as triangular fuzzy sets (Figure S2). The definitions were chosen like this to avoid extremely low/high values being labelled as medium: values below 25 or above 75 are not considered medium. This limits the number of rules that are triggered and allows for more extreme output values (i.e. values closer to the minimum and maximum of the domain). Experiments have shown that the difference is not big, and that the issue can be overcome also by using more fuzzy sets. A similar approach is performed for the determination of fuzzy sets for the auxiliary grid.

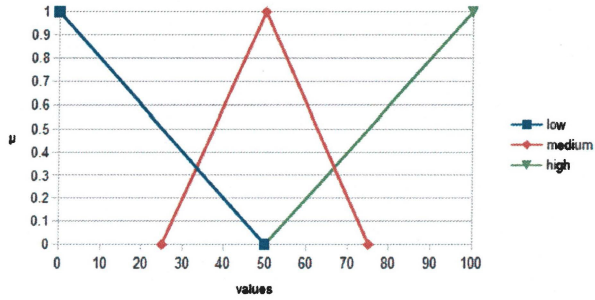


Figure S2. Example of the fuzzy sets used to define low, medium and high values for both input and auxiliary grid.

The value outputted by the inference system is however interpreted as a weight. Its value will be rescaled so that all the sum of all output cells that overlap a given input cell still has the same value. As such, the absolute value of the output is not important, but it is chosen to be a value in the range 0-100. The sets are defined as illustrated on Figure S3.

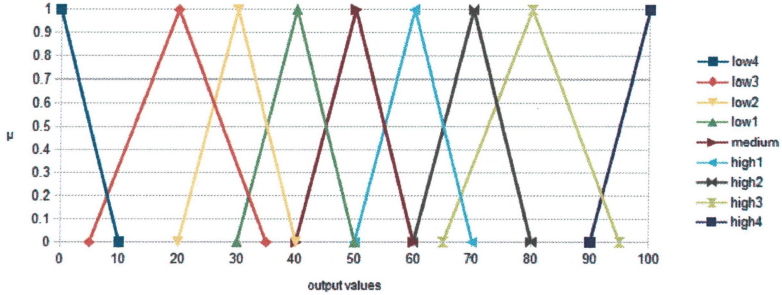


Figure S3. The sets to define the labels of the output values of the inference system.

The fuzzy sets are distributed symmetrically, but are not evenly spaced. This was done achieve a more emphasized impact of really low or really high values. A value below 5 is only considered to be *low4*, a value between 5 and 20 is not considered to be greater than *low3*. This makes sure the extreme values are still considered to be quite extreme, preventing the rulebase from averaging the values too much. At present, the output is returned as a relative value, which is then rescaled to make the output correctly reflect the input. This is not as accurate as it could be; work is in progress to have the rulebase return final values, but the current approach is sufficient to verify the workings of the methodology. To convert the outputted fuzzy set to a crisp set, the centre of gravity of the fuzzy set is considered.