Raport Badawczy
Research Report

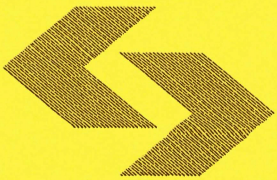
RB/9/2014

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

Solving the map overlay problem with a fuzzy approach

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.

Keywords: map overlay, gridded data, fuzzy processing.

1. Introduction

In order to compare different countries, the FCCC requires a single national value per country for e.g. CO2 emissions that stem from fossil-fuel burning. The authors in Boychuk and Bun (this issue); Jonas and Nilsson (2007) explain that for countries with good emission statistics, the national fossil-fuel CO2 emissions are believed to exhibit a relative uncertainty of about +/- 5% (95% CI), but that a sub-national approach can differ considerably (i.e., the +/- 5 % for the 95% CI does not hold any more). This is due to uncertainty at various levels, both uncertainty inherently present in the data, but also uncertainty introduced by processing and pre-processing the data. The International Workshop Series on Uncertainty in GHG Emission Inventories focuses on both the presence and on techniques on understanding, modelling and decreasing these uncertainties (Bun et al, 2007; Bun et al., 2010, Jonas et al., 2010). The sub-national data are usually obtained through the analysis of data coming from different sources, processed and combined into a national value, but the way the data are processed has a big impact on the introduced uncertainty and consequently on the results. To eliminate the inter-country uncertainty, uniform and well-tested methodology should be used when building on sub-national emission approaches. However, even when using the same methodology, the uncertainty introduced by the processing of data is dependent on the source formats of the data, and how it behaves under subsequent processing. The solution to this is either obtaining data in a more similar way, to make all data compatible, but with most infrastructures in place, changes to this are unlikely to happen. The alternative is to pre-process the data so that the data exhibit a similar behaviour under the subsequent processing. The data are often represented in a gridded format, yet often different grids (e.g. CO_2 emissions and land use) are incompatible. By transforming them to compatible, matching grids, the processing should yield more consistent results.

This article presents a novel approach to pre-process data, to transform the data (e.g. emission data) so that it is better suited to be combined with other data (e.g. land use) while at the same time keeping the uncertainty and errors introduced by this transformation to a minimum. As such, this methodology is at a very low level in the processing chain, but any decrease in uncertainty at such a low level should provide far more reliable results at the end of the processing and thus allow for more accurate analysis and comparison.

Commonly, data relating to different topics come from different sources: land use data can be provided by one source, emission data from another source, population data is again obtained elsewhere. Usually, the data are provided in a gridded format (Rigaux et al., 2002, Shekhar and Chawla, 2003), which means that the map (or the region of interest) is overlayed with a grid dividing the map in different cells. In the case of a rectangular grid, each grid cell will be a rectangle or a square. With each cell, a numeric value is associated; which is deemed to be representative for the cell. The cell is however the smallest item for which there is data: the value associated with the cell can be the accumulation of data of 100 different points in the cell, or one single point in the cell, several line sources, ...; there is no difference in appearance between these cells and no way of knowing this once the data is presented in the grid format. This is illustrated on Figure 1(a). There are however several problems with the data, particularly when the data need to be combined. As the data are obtained from different sources, the format in which they are provided can differ: the grids could not line up properly, the size of the grid cells can be different, the grids could be rotated compared to one another, etc; as illustrated on Figure 1(b)-(e). This makes it difficult to relate data that is on different grids to each other and thus introduces uncertainty or errors. Additionally, not all data is complete, and cells of the grid may be without data.

This article does not concern any specific data analysis, nor conclusions that directly relate to climate change, but it introduces a novel solution method to transform a grid to match a different grid, a process that is used in many climate related studies. The proposed methodology makes use of data analysis, geometric matching and mathematical connections to solve format mismatches, it does not consider higher concepts that relate to the meaning of the data (e.g. using ontologies to match differently labelled data, as in (Duckham and Worboys, 2005)). The proposed methodology is still in very early stage, and as such the examples shown are quite simple but prototype implementation and experiments on small samples show promising results. The methodology is expected to be able to cope with uncertainties and missing information, but focus at this moment is on developing the basic workings of the method. In Section 2, the map overlay problem along with current solution methods will be described. A reasoning about the problem and the possible use of added knowledge is also covered here. Section 3 considers how the intuitive approach can be simulated using techniques from artificial intelligence; it briefly introduces the necessary concepts (fuzzy set theory, fuzzy inference system) that will be used further. and describes the new methodology. In Section 4, the methodology is applied on some examples and directions for future research are listed. Section 5 summarizes the article and contains conclusions.

2. The map overlay problem

2.1 Problem description

Spatially correlated numerical data are often represented by means of a data grid. This grid basically divides the map (or the region of interest) in a number of cells. These cells are commonly equal in size and shape (regular grid) and most commonly are rectangles or squares. Each cell is considered to be atomic in the sense that it is not divided in smaller

parts, and contains aggregated information for the area covered by the cell. With each cell, a numeric value is associated that is deemed representative for the cell. If we consider the example of the presence of a greenhouse gas, then the value associated with the cell indicates the amount of this particular gas in that particular cell. In reality, this amount may be evenly spread over the entire cell or it may be concentrated in a very small part of the cell; but as the cells are the smallest object considered, there is no way to differentiate between them. This is illustrated on Figure 1(a).

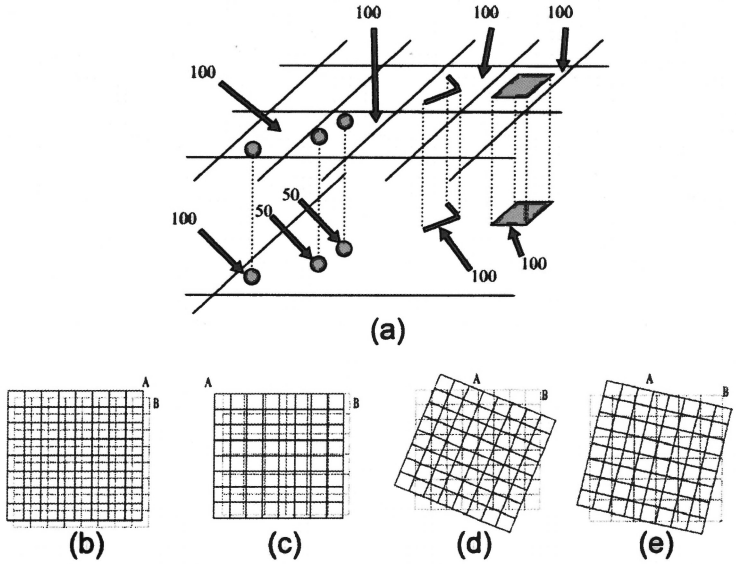


Figure 1. Different data distributions within a grid cell that result in the same value for the grid cell are shown in (a). The examples are: a single point source of value 100, two point sources of value 50, a line source of value 100 and an area source of value 100. Each of these are such that they are in one gridcell, which then has the value 100. When viewing the gridcell, it is not known what the underlying distribution is. Different incompatible grids are shown in (b)-(e): a relative shift (b), a different gridsize (c), a different orientation (d) and a combination (e).

Commonly, data from different sources need to be combined to draw conclusions: for instance, relating the measured concentrations of a particular gas in the atmosphere to the land use would require data from both concentrations and land use. While both data can be represented as gridded data, the grids used often don't match: not only can there be a difference in cell sizes and shapes, but one grid can also be rotated compared to the other grid, translated, or any combination of these. This is a common problem with many data in literature, called the *map overlay problem*; the data are then said to be *incompatible*; some examples are shown on Figure 1(b)-(e).

To make these different datasets compatible, it is necessary to transform one of the grids to match the layout of the other grid: it needs to have the same number of grid cells, oriented in the same way, so that there is a 1:1 mapping of each cell in one grid to a cell in the other grid. The map overlay problem concerns finding a such a mapping: it considers an

input grid which contains data and an target grid that provides the new grid structure on which the input grid needs to be mapped. As mentioned before, nothing is known at a scale smaller than the cells; which makes the mapping one grid to another a very difficult problem. Consider the simple example on Figure 2(a).

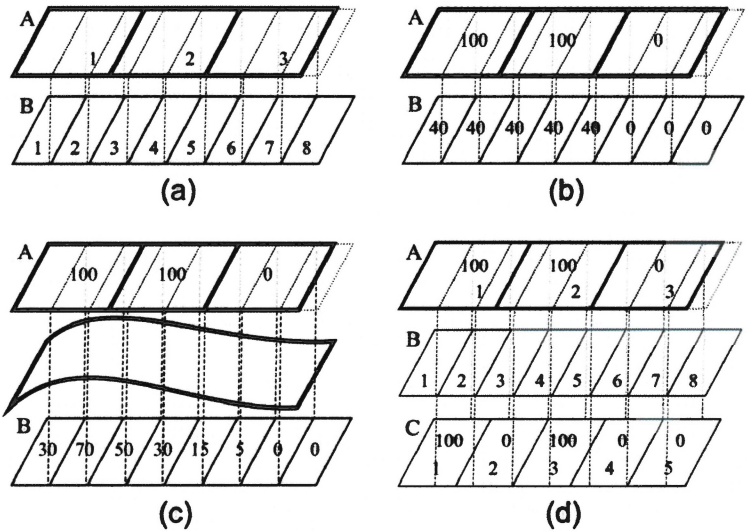


Figure 2. Examples to explain the problem: (a) Problem illustration: remapping grid *A* onto grid *B*, (b) Areal weighting: the value of each output cell is determined by the amount of overlap, (c) Areal smoothing: the value of each output cell is determined resampling a smooth surface that is *fitted* over the input data, (d) Intelligent reasoning using additional data: grid *C* supplies information on the distribution, which can be used to determine values in the output grid.

Remapping the values of the gridcells of the inputgrid A to the output grid B is done by determining the values of x_i^j in these formulas:

$$f(B_1) = x_1^1 f(A_1)$$
$$f(B_2) = x_2^1 f(A_1)$$
$$f(B_3) = x_3^1 f(A_1) + x_3^2 f(A_2)$$
$$= (1 - x_1^1 - x_2^1) f(A_1) + (1 - x_4^2 - x_5^2) f(A_2)$$
$$f(B_4) = x_4^2 f(A_2)$$
$$f(B_5) = x_5^2 f(A_2)$$

Where in x_i^j, the index i refers to the cell number in the output grid, and j refers to the cell number in the input grid. This can be generalized as:

$$f(B_i) = \sum_j x_i^j f(A_j) = \sum_{j \mid A_j \cap B_i \neq \emptyset} x_i^j f(A_j)$$

with constraints that

$$\forall j, \sum_i x_i^j = 1$$

While this looks straight forward, the problem is finding the values of x_i^j. In more complicated examples, it is obvious that more than two grid cells can contribute to the value of a new grid cell. From the example on Figure 2(a), it can also be seen that there is no single solution: there are different possible values for B_1 and B_2, as long as their sum is constant.

The key to resampling the original grid to the new grid, is figuring out the true distribution of the data; thus going into further detail than the grid cells offer. Most current solution methods either assume a distribution of the data or aim to estimate the distribution of the data, and resample in order to match the new grid. The output grid is specified by the user; the initial map overlay problem concerns transforming the input grid to this grid. The proposed method uses an additional grid, also specified by the user, to help transform the input grid, but the grid specification (cell size, orientation) does not have to match either input or output and does not change the output format. The additional grid should contain data that has a known and established correlation to the input data. The grid on which the data is provided does not have to match either input or output grids; however, the finer the grid, the better the results for transforming the input grid.

2.2 Current solution methods

Transforming one grid to another grid is comparable to determining what the first grid would be if it were represented by using the same grid cells as the second grid. In literature, there exists a number of solution methods. The gridcells as shown in Figure 2(a) will be used to explain the most common methods, for a more detailed overview we refer to (Gotway, 2002). In the example used, the input grid A contains three cells, the output grid B contains eight cells.

2.2.1 Areal weighting

The simplest and most commonly used method is areal weighting. It uses the portion of overlap of the grid cell to determine what portion of its associated numerical value will be considered in the new grid. This approach is very easy and straightforward. The result of areal weighting is illustrated on Figure 2(b). This means that the values of x_i^j are determined by the surface area S:

$$f(B_1) = S(B_1) f(A_1)$$
$$f(B_2) = S(B_2) f(A_1)$$
$$f(B_3) = S(B_3 \cap A_1) f(A_1) + S(B_3 \cap A_2) f(A_2)$$
$$f(B_4) = S(B_4) f(A_2)$$
$$\ldots$$

Basically, in this approach, it is assumed that the data in the cell are evenly distributed throughout the cell and that all cells are considered to be completely independent of one another. Resampling can be done over any grid without difficulties. In some situations, this may indeed be the case or at least a close enough assumption to justify using it. The spread of a gas in the atmosphere in the absence of extreme sources is an example of this. However, when the associated numeric data is the result of a small number of extreme sources in the grid cell (e.g. a factory), then this approach may lead to either an over- or an underestimate in the new grid cell, depending on whether or not the factory is in the overlapping area.

2.2.2 Spatial smoothing

Spatial smoothing is a more complicated approach than areal weighting. Rather than assume that the data is evenly spread out over a cell, the distribution of the data within a cell is dependent on neighbouring cells.

This is achieved by considering the grid in three dimensions, with the third dimension representing the associated data. In spatial smoothing methods, a smooth three-dimensional surface is fitted over this grid, as illustrated on Figure 2(c), after which the smooth surfaced is sampled using the target grid. This method therefore does not assume that the data modelled by the grid is evenly distributed over each cell, but assumes a smooth distribution over the region of interest: if the value of a cell is high, one expects higher values closer to it in the surrounding cells. In many situations, this method is more accurate than the previous method, but is still unable to cope with data that in reality is concentrated in a small area of the cell. This is for instance the case when modelling air pollution, and a single factory is responsible for the value that will be associated with the grid cell in which it is contained (a point source): the presence of a point source in one cell does not imply sources close to it in neighbouring cells (in some urban planning schemes, it might even be the opposite, to avoid placing too many point sources too close to each other).

2.2.3 Regression methods

In regression methods, a relation between both grids is examined, and patterns of overlap are established. Different methods exists, based on the way the patterns are established. (Flowerdew and Green, 1994) determine zones, which are then used to establish a relation. This is then combined with an assumption of the distribution of the data (e.g. Poisson) in order to determine values for the incompatible zones. Several underlying theoretical models can be used, but all regression methods require key assumptions that normally are not part of the data and cannot be verified using the data. These assumptions mainly concern the distribution of the data, e.g. if the data is distributed in a Poisson or binomial distribution.

2.3 Using additional knowledge

2.3.1 Data fusion

The described problem to some extent resembles the problem described in (Duckham and Worboys, 2005). Both the problem and solution are however is completely different: the authors in (Duckham and Worboys, 2005) combine different datasets that relate to the same area of interest in order to create a new dataset that has the combined information of both source data sets. This combined information can be richer or have a higher accuracy. Their approach however is not intended for numerical data, but for labelled information. The different datasets can use a different schema (set of labels) to describe regions in the region of interest (the example uses land coverage and land use terms). As the labels are not always fully compatible, the authors propose a method of linking both schemas with a common ontology, and obtaining geometric intersections if the labelled regions do not match. The authors in (Fritz and Lee, 2005) tackle the data fusion problem using a different approach. An expert supplies input regarding the assigned labels in different datasets; this knowledge is then modelled and matched using a fuzzy agreement. This allows the labels in different sets to be compared and combined properly.

While both these approach are also using multiple datasets, the type of data processed and the goal of the processing is quite different from what is presented in this article. In the aforementioned data fusion approaches, the goal is the combine annotations and labels added

in different datasets by different people. This is not numerical information, but e.g. true land use information. The datasets are also not represented by grids but by vectorial maps.

2.3.2 Intuitive approach to grid remapping

The methods mentioned in Section 2.2 work on gridded data and transform the grid without any possibility to use additional data that might be available, and where some key assumptions regarding data distribution are implied within the methods. While at first it seems that the only knowledge available is the input grid, it is very likely that there is additional knowledge. Consider for instance an input grid that represents CO_2 concentrations on a course grid. From other research, the correlation between CO_2 levels and traffic is known. This means that we can use this known correlation to improve the CO_2 data set we have by using traffic information that is also available for the same region. Of course, the correlation between the input and additional datasets should be known beforehand, as this is a key assumption of the method. When this correlation is known, this information can be used to transform the grid that represents CO_2 emissions to a grid with different cell size or with different orientation.

The presented method allows for additional data to be taken into account when resampling the data to a new grid. Suppose additional data, which relates to the data in the input grid is available in a grid containing 5 cells as shown on figure 5.

Based on the values in the additional grid *C*, it is possible to guide the distribution of the values modelled on grid *A* to the new grid *B*. A low value in a cell of grid C suggests that the values in the overlapping cells of the output grid should also be lower.

By strictly interpreting this additional grid, it is possible to intuitively derive a simple distribution: proportional values for $f(B_1)$, $f(B_2)$, $f(B_4)$, $f(B_5)$. The strict interpretation means that the cells in the output grid B should have a value that is proportional to both the input grid A and the auxiliary grid C. In many situations however, this cannot be achieved, as data in grids can be slightly contradictory, a consequence of the fact that grids are approximations of the real situation. It is often not possible to derive a distribution when interpreting the related data grid too strict; but it is possible to derive a distribution that is still consistent with the input grid, and to some extent follows the related grid *C*. The related grid is thus only used to help determine the original, unknown distribution. Obviously, the grid used to help in transforming the data should contain a well established known relation to the input data. If the relationship between input grid and auxiliary grid are under investigation, any usage of the auxiliary grid in transforming the input grid may distort conclusions on the relationship between both grids.

To come to an intuitive solution consider cell B_3. To derive the $f(B_3)$, it is necessary to look at the grids *A* and *C* in the area around B_3. The cells that are of interest are A_1 and A_2. Both have the same value, so they will not provide much information. On the other hand, the cell C_2 that overlaps B_3 has a very low value ($f(C_2) = 0$), whereas its neighbouring cells have high values ($f(C_1) = f(C_3) = 100$). As the data in grid *C* are known to be related to the data in *A*, we can conclude that the data of grid A for this region should be more spread towards the neighbouring cells of B_3, so the value $f(B_3)$ should be low.

Consider B_4. Again, the values of the overlapping cells in *A* are the same, so this will not influence the result. But the overlapping cell of grid *C*, C_3 has a high value. The neighbouring cells of C_3 have a low value. This basically implies that the distribution of *A* over the cells in *B* should also be lower in the proximity of cell C_3.

Finally, consider B_5. Here, the values of the overlapping cells in A are different: $f(A_2)=100$, $f(A_3)=0$. The distribution of the data in grid C tell us that in B_5, the value should be lower: no contribution from A_3, and C_3 has a much higher value than C_4.

The use of additional information, in this example a single grid, can for sure contribute to obtaining a distribution that is still consistent with the input grid, but at the same time takes into account added available knowledge. From the examples though, it can be seen that it is not always possible to find a unique solution, implying there is still some uncertainty on the accuracy of the newly obtained grid.

The above example only uses a proportional or inverse proportional relationship between cells. This relationship is however only considered at a local scale, meaning that high and low for both input grid and additional grid are defined for the location under consideration, independent of the definition of other locations. As such, the connection between the input grid and the additional grid is not quantitatively verified, but only relative values are considered which makes the approach not dependent on linearity or non-linearity. By considering different rules, it is even possible to model different connections, e.g. : *if a value is high or a value is low, then the output value should be high.* The ultimate goal is to allow multiple additional data layers, and allow for different possible combination (e.g. high value in one and low value in another can yield a result that might well be the same as low value in one and high value in the other. On the other hand, also considering the more quantitative connection between the layers, can also provide better results. Both these aspects are part of future research.

3. Using intelligent techniques

3.1 Introduction to fuzzy sets and fuzzy inference

3.1.1 Fuzzy sets

Fuzzy set theory was introduced by Zadeh in (Zadeh, 1965) as an extension of classical set theory. In a classic set theory, an object either belongs to a set or it does not belong the set. In fuzzy set theory, the objects are assigned a membership grade in the range [0,1] to express the relation of the object to the set. These membership grades can have different interpretations (Dubois and Prade, 1999): a veristic interpretation means that all the objects belong to some extent to the set, with the membership grade indicating the extent; whereas a possibilistic interpretation means there is doubt on which elements belong, now the membership grade is expressing the possibility that an element belongs to the set. Lastly, it is also possible for the membership grades to represent degrees of truth. In (Dubois and Prade, 1999) it was shown that all other interpretations can be traced back to one of these three. The formal definition of a fuzzy set $\tilde{A}$ in a universe U is given below

$$\tilde{A}=\{(x,\mu_{\tilde{A}}(x)\mid x\in U)\}$$

Its membership function $\mu_{\tilde{A}}(x)$ is

$$\mu_{\tilde{A}}:U\rightarrow[0,1]$$
$$x\mapsto\mu_{\tilde{A}}(x)$$

Various operations on fuzzy sets are possible: intersection and union are defined by means of functions that work on the membership grades, called respectively t-norms and t-conorms. Any function that satisfies specific criteria is a t-norm, respectively t-conorm and can be used to calculate intersection or union (Klir and Yuan, 1995; Zimmerman, 1999). Commonly used t-norms and t-conorms are the Zadeh-min-max norms, which use minimum

as the intersection and the maximum as the union (other examples are limited sum and product, Lukasiewicz norm, ...).

Fuzzy sets can be defined over any domain, but of particular interest here are fuzzy sets over the numerical domain, called fuzzy numbers: the membership function represents uncertainty about a numeric value. The fuzzy set must be convex and normalized (some authors also claim the support must be bounded, but this property is not strictly necessary) (Klir and Yuan, 1995). Using Zadeh's extension principle (Zadeh, 1965), it is possible to define mathematical operators on such fuzzy numbers (addition, multiplication, etc.). Fuzzy sets can also be used to represent linguistic terms, such as "high", "low"; this allows one to determine which numbers are considered high in a given context. Linguistic modifiers also exist and are usually a function that alters the membership function for the term it is associated with, allowing for an interpretation of the words like "very" and Somewhat".

Finally, it is necessary to make a distinction between an inclusive and an exclusive interpretation: are values that match "very high" still considered to be "high"? In real world, people could say about a person: "he is not tall, he is very tall", which is an exclusive interpretation: "very tall" does not imply "tall". The main difficulty when using fuzzy sets is the definition of the membership functions: why are the fuzzy sets and membership grades chosen as they are, and on what information this choice is based.

3.1.2 Fuzzy inference system

A fuzzy inference system is a system that uses a rulebase and fuzzy set theory to come to solutions for given (numeric) problems (Mendel, 2001; Klir and Yuan, 1995). The rulebase consists of fuzzy premises and conclusions; it is comprised a set of rules that are of the form

$$\text{if } \underbrace{x \text{ is } A}_{\text{premise}}, \text{ then } \underbrace{y \text{ is } B}_{\text{conclusion}}$$

Here "x is A" is the premise and "y is B" is the conclusion; x and y are values, with x the input value and y the output value. Both are commonly represented by fuzzy sets, even though x usually is a crisp value (crisp means not fuzzy). In the rule, A and B are labels, such as "high" or "low", also represented by fuzzy sets as described above.

The "is" in the premise of the rules is a fuzzy match: this will return a value indicating how well the value x matches with label A. As all the rules are evaluated and the values are fuzzy, it is typical that more than one rule can match: a value x can be classified as "high" to some extent and at the same time as "low" to much lesser extent. All the rules that match will play a part in determining the outcome, but of course the lower the extent to which a rule matches, the less important its contribution will be. It is possible to combine premises using logical operators (and, or, xor) to yield more complex rules. As multiple rules match, y should be assigned multiple values by different rules: all these values are aggregated using a fuzzy aggregator to yield one single fuzzy value. For each rule, the extent to which the premise matches impacts the value that is assigned to y.

The "is" in the conclusion is a basic assignment and will assign y with a fuzzy set that matches the label B. It is important to note that x and y can be from totally different domains, a classic example from fuzzy control is "if temperature is high, then cooling fan speed is high".

While the output of the inference system is a fuzzy set, in practise the output will be used to make a decision and as such needs to a crisp value. To derive a crisp value (defuzzification), different operators exist. The centroid calculation is the most commonly used; it returns the centre of the area under the membership function.

3.2 Defining the inference system

The parameters for the used inference system are derived from generated sample cases, for which an optimal solution is known. The relations between the found parameters and the optimal solution are then reflected in the created rules. Due to the more technical nature of this explanation, and the strict page limitation, the detailed explanation of the procedure is available as supplementary material in (Online Resource 1).

Appendix 1 contains the details of how the inference system is defined. Appendix 2 goes into further details of the parameters, while Appendix 3 covers the automatic selection of the best parameters for a given input.

4. Experiments

4.1 Description of the results

In this section, results of the methodology applied on the example in Figure 2(d) for several inputs will be shown and discussed. The input grid A has three grid cells, the output grid B contains eight grid cells, but does not exactly overlap with A and the auxiliary grid C has five grid cells and overlaps full with grid A. It is obvious that the grids are not aligned properly.

Table 1 holds the data for the inputgrid, the auxiliary grid; and the computed output grid. The first three cases have a distribution of the input data such that $f(A_1)=f(A_2)=100$ and $f(A_3)=0$; the last three cases have the input distribution such that $f(A_1)=f(A_3)=100$ and $f(A_2)=0$. In each of the cases, a different distribution of the auxiliary grid was considered, as shown on Table 1. Figure 3 offers a graphic view of each of the cases. For each case, the grid cells are shown; the surface area of the circles is representative for the values associated with the grid cells. Consequently, the sum of the areas of the circles in grid B equals the sum of the areas in grid A. There is no quantitative relation assumed between the auxiliary grid B and the grid A, it is just assumed that high values in B are an indication for high values in A.

Table 1. Overview of the cases used in the simulations. The grid layout is illustrated on Figure 2(d), these results are graphically illustrated on Figure 3.

	input grid A			auxiliary grid C					output grid B							
case 1	100	100	0	100	100	100	0	0	25	50	52	27	46	0	0	0
case 2	100	100	0	100	100	0	0	0	25	50	57	23	46	0	0	0
case 3	100	100	0	100	0	0	100	0	29	50	44	23	54	0	0	0
case 4	100	0	100	100	0	0	0	100	29	50	21	0	0	26	37	37
case 5	100	0	100	100	100	0	0	100	25	50	25	0	0	26	37	37
case 6	100	0	100	0	100	0	0	100	21	50	29	0	0	26	37	37

The behaviour of the methodology under the different cases is clearly illustrated in Figure 3.

Consider output cell B_2. From (Online Resource 1), the considered cells that play a part are: A_1 (proportionally), C_1 and C_2 (proportionally), A_2 (inverse proportionally). In all six cases, the value assigned to B_2 is the same. The reason is that, while the values of the involved cells differ (C_2 changes value), the locally computed definition for the fuzzy sets defining low and high for the auxiliary grid also modifies. They modify in such a way, that the differing input is cancelled out. In the system, there is no difference between a value of for example 100 when low is defined as 0 and high is defined as 100, and a value of for example 200 when low is defined as 0 and high is defined as 200.

Cell B_3 shows a much bigger variation. The cells involved in determining the output value are A_1 and A_2 (proportionally), C_2 (proportionally), C_1 and C_3 (inverse proportionally). In case 1, the proportional and inverse proportional data almost cancels out, resulting in a value close to 50. In case 2, the value for C_3 is much smaller than in case 1, which results in the larger value of 57 for B_3. In case 3, the value of C_2 is decreased, and as it is has a proportional relation to B_3, the value for B_3 is again decreased. It is smaller than in case 1, as the higher value of C_3 causes a higher value in B_5, which compensates. The difference in cases 4 to 6 are explained by the change in the proportional and inverse proportional data from the auxiliary grid: in case 4 there is more inverse proportional than proportional (the value of C_1 is greater than the value of C_2), in case 5 they are equal, and in case 6 the proportional value is greater than the inverse proportional.

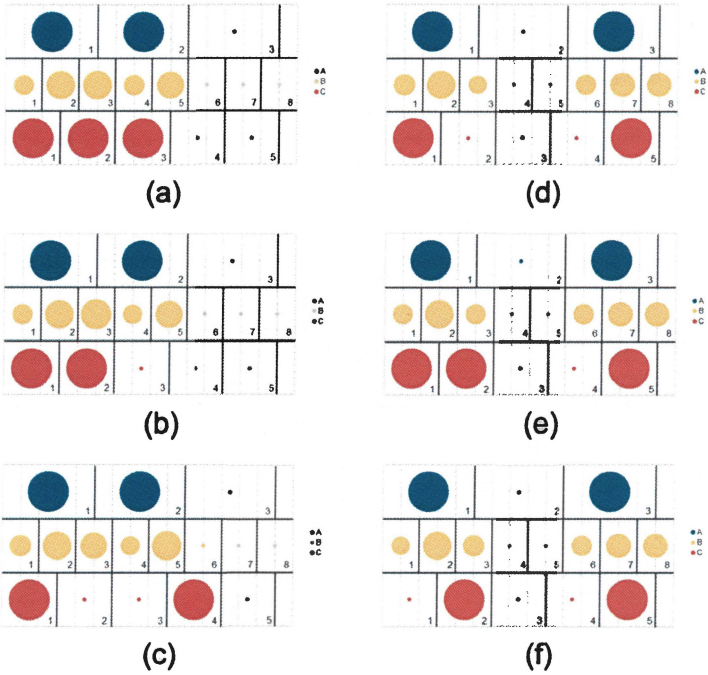


Figure 3. Illustrations for the different cases from Table 1: *A* is the input grid, *B* is the output grid and *C* the auxiliary grid. The gridcells are drawn above each other for visibility purposes, but should cover each other as shown on Figure 2(d). The size of the circles reflects the relative value of the associated cell (a small circle is shown for 0 values, for illustration purposes).

Cell B_4 is influenced by A_2 (proportional), C_3 (proportional), C_2 and C_4 (inverse proportional). The first case have a greater value than cases 2 and 3, as C_3 has a much higher value. The second and third cases result in the same value, as the change in the auxiliary grid also changes the definition for high, causing the change in values to be nullified. The value 0 in the last 3 cases is due to the overlapping input field having 0 as associated value.

Cell B_5 is determined by A_2 (proportional), C_3 and C_4 (proportional), A_3 (inverse proportional) and C_2. The first two cases are the same, as the definitions for high for the

auxiliary grid is changed also. Case 3 shows a higher value, as there is more proportional contribution from C_4.

The cells B_6, B_7 and B_8 can be considered together. They all are 0 in the first 3 cases, as the overlapping input field has 0. In cases 4,5 and 6, the latter two have higher values, which is the expected behaviour due to the values of C_4 and C_5.

4.2 Observations of the methodology

From the cases on Figure 3, it can be seen that the goal of using an auxiliary grid to guide the new distribution yields some interesting results. In general, the methodology does not yield contradictory effects: the output grid fully complies with the input grid. Compared to the traditional approaches (e.g. areal weighting, which would provide the same result for all first three cases, and the same result for all last three cases, it is clear that the additional data has an effect on the result.

The distribution in the output grid to some extent follows the auxiliary grid, but there are some exceptions. In the last three cases, the results look consistent and as expected: larger values where the auxiliary grid overlaps, smaller values elsewhere. In the first two cases, the larger value of cell B_5 stands out. This is mainly explained by the fact that B_5 fully overlaps with A_2, and by the fact that the values of cells considered in the auxiliary grid cancel each other out, or the definition of high for the auxiliary values changes to yield this effect. Similarly, the value of B_3 in case 3 stands out as counter intuitive, but with a value of 44 it still is quite a lot smaller than in cases 1 (52) and 2 (57), which is consistent with the desired result. A similar observation can be made for cell B_6 in the last three cases: its value is perhaps higher than would be desired based on the auxiliary grid, but still the values for the cells that overlap with the cells of the auxiliary grid that have higher values also have higher values than B_6.

The results appear to reach the desired goals, but still further testing and development of the methodology is required.

4.3 Future developments

The presented approach is a first concept and prototype implementation of a new methodology that shows promising results, and as such justifies further research. The prototype allows us to experiment and see how the system behaves and derive why it behaves like that. The outcome of the fuzzy inference system is dependent on a large number of parameters.

First, there are the parameters that concern the geometrical aspects of the problems, the definition of the cells that are considered to have an influence on a given output cell. This not only concerns choosing which cells will take part in determining the value for the output cell, but also determining the behaviour (proportional or inverse proportional) and adding weights to the cells to decrease their influence (e.g. if the distance becomes too great to be relevant). In the current implementation, no quantitative relationship between auxiliary and input grid is assumed, meaning that the quantitative relation between input grid and additional grid is only considered for the vicinity of a cell. A quantitative relation on a bigger scale can however be used to derive how big the impact of the auxiliary cell should be, and as such should provide better results. This is however not a trivial step, as too tight a relationship may cause too narrow constraints and consequently prevent the system from reaching a good solution when data is contradictory or missing.

Second, there are the parameters that define the rulebase: this is not only the number of rules, but also the definition of the rules themselves and weights assigned to the rules. At present, the number of rules was derived from all possible combinations of values of cells that play a part. It is however possible to limit the rules and for instance consider a fixed

number of rules that are determined automatically by means of training data. The full impact of this is quite difficult to estimate at this time though. Each of the rules can also be assigned a weight, and at present, lower weights were assigned to contradicting rules. In the examples in this article, this yielded little impact, as the data used in the input was not really contradictory. This parameter may become more important when confronting the system with real world data and missing data.

Last, there are the parameters that relate to the fuzzy sets used and their definitions. This includes the definitions of the fuzzy sets that represent high, low; the definitions of minimal and maximal values that are used in these fuzzy sets and the number of sets that are considered for both input and output. Also, the definitions of minimum and maximum of the domain (explained in (Online Resource 1)) can be improved: the current definitions limit possible input sets too much.

Appendix 4 expands on the experiments presented here, and uses bigger, gridded data that are processed using the presented technique and the quality of the processing is evaluated.

5. Conclusion

In this article, a completely novel approach to the map overlay problem was presented. The described methodology is in very early stage, but shows interesting results. Rather than assuming a distribution of the data, knowledge from external data that are known to relate to the input data are used to find a more optimal distribution of the data after transformation to a new grid. The methodology uses concepts from fuzzy set theory and algorithms from artificial intelligence in order to mimic reasoning about the input data. The results show that in simple cases, the methodology achieves the pre-set goal, but additional testing and fine tuning is necessary. The current prototype should allow for the processing of larger datasets. The next step is to generate artificial but large scale data, in order to fine tune the workings of the methodology using a fully controlled environment and to assess the performance; both in accuracy and processing speed.

References

Boychuk, K., Bun, R.: Regional spatial cadastres of ghg emissions in energy sector: Accounting for uncertainty. Climatic Change current volume (20xx)

Jonas, M., Nilsson, S.: Prior to economic treatment of emissions and their uncertainties under the kyoto protocol: Scientific uncertainties that must be kept in mind. Water, Air, & Soil Pollution: Focus 7(4-5), 495–511 (2007)

Bun, R., Gusti, M., Kujii, L., Tokar, O., Tsybrivskyy, Y., Bun, A.: Spatial ghg inventory: Analysis of uncertainty sources. a case study for ukraine. Water, Air, & Soil Pollution: Focus 7(4-5), 483–494 (2007)

Bun, R., Hamal, K., Gusti, M., Bun, A.: Spatial ghg inventory on regional level: Accounting for uncertainty. Climatic Change 103, 227–244 (2010)

Jonas, M., Marland, G., Winiwarter, W., White, T., Nahorski, Z., Bun, R., Nilsson, S.: Benefits of dealing with uncertainty in greenhouse gas inventories: Introduction. Climatic Change 103, 3–18 (2010)

Rigaux, P., Scholl, M., Voisard, A.: Spatial databases with applications to GIS. Morgan Kaufman Publishers (2002)

Shekhar, S., Chawla, S.: Spatial databases: a tour. Pearson Educations (2003)

Duckham, M., Worboys, M.: An algebraic approach to automated information fusion. International Journal of Geographic Information Systems 19(5), 537–558 (2005)

Gotway, C.A., Young, L.J.: Combining incompatible spatial data. Journal of the American Statistical Association 97(458), 632–648 (2002)

Flowerdew, R., Green, M.: Spatial analysis and GIS; eds. Foterhingham S. and Rogerson P., chap. Areal interpolation and types of data, pp. 141–152. Taylor & Francis (1994)
Fritz, S., See, L.: Comparison of land cover maps using fuzzy agreement. International Journal of Geographic Information Science 19, 787–807 (2005)

Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)

Dubois, D., Prade, H.: The three semantics of fuzzy sets. Fuzzy Sets and Systems 90, 141–150 (1999)

Klir, G.J., Yuan, B.: Fuzzy sets and fuzzy logic: theory and applications. Prentice Hall, New Jersey (1995)

Zimmerman, H.J.: Practical applications of fuzzy technologies. Kluwer Academic Publishers (1999)

Mendel, J.M.: Uncertain rule-based fuzzy logic systems, Introduction and new directions. Prentice Hall (2001)

Online Resource 1: ESM1.pdf: Supplementary material containing the description of the creation of the fuzzy rulebase system.