Using a fuzzy inference system for the map overlay problem

Dr. Verstraete Jörg[1]

[1] Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, Warsaw, 01-447, Warsaw
jorg.verstraete@ibspan.waw.pl
Department for Telecommunications and Information processing, Ghent University
Sint Pietersnieuwstraat 41, Gent, 9000, Belgium
jorg.verstraete@telin.ugent.be

Abstract

In this contribution, the problem of ill aligned spatial data is considered. The problem is commonly known as the map overlay problem, and occurs when data from different grids are combined. When the different grids don't line up properly, determining what portion of the data associated with a tile in one grid is relevant for the data associated with a partially overlapping tile in the latter grid becomes a problem. As it is in general not possible to derive an exact value, we opted for an approach that results in fuzzy sets. Rather than process the data itself, a seemingly intelligent system to make a decision on which portions could be relevant, was developed. The presented approach makes use of a fuzzy inference system, a system built up of a number of if-then rules containing fuzzy predicates. These rules are used to evaluate set of input values to yield one or more output values. The input values can be fuzzy sets themselves, output values are always fuzzy sets (in our case fuzzy numbers) which are then defuzzified.

Keywords: Map overlay problem, fuzzy inference system

1. Introduction

When dealing with gridded data, it can be necessary to combine data from different sources: one source could supply data of emissions of specific gasses, whereas another source could supply land use information; the combination of both is needed to derive a link between the data. In our case, one grid concerns emission data, while a finer grid contains covariate data, e.g. information of land use or population data that is known to have a relation with the emissions represented in the former grid. In literature, there have been a number of approaches, ranging from simple aerial weighting to spatial smoothing and various regression methods to solve this problem [1]. In general however, it can be concluded that no exact solution is possible: the gridded data itself is usually an approximation, interpreting it to match a different grid will only increase the uncertainty or the imprecision. Unlike the aforementioned methods that describe algorithms to manipulate the data to better match the other grid, we considered a different approach, using fuzzy set theory and a fuzzy inference system.

The more detailed workings of fuzzy sets and the inference system will be explained in the next sections. The concept of the approach is that we derived rules that describe how data of one grid should be redistributed over the second grid. These rules were mainly derived from specific example cases; once implemented, the inference system applies these rules on the real data and redistributes the values accordingly. The accuracy of the result depends largely on the number of rules considered, the fuzzy sets used to represent the data and on how well the rules reflect the desired behaviour.

In the next section, we will introduce the fuzzy inference system. For this, a brief introduction in fuzzy set theory is required, with some explanation on fuzzy numbers and representation of linguistic terms. After this, the concept of the fuzzy rulebase and its workings can be explained. The subsequent section will elaborate on the application of the inference system in the context of the map overlay problem. First, a simple example will be used to derive the rules and explain the concept. A more advanced example will then illustrate the feasibility of this approach so far. The conclusion will summarize the findings.

2. Fuzzy inference system

2.1. Introduction to fuzzy set theory

Fuzzy set theory was introduced by Zadeh in [2] as an extension of classical set theory. In a fuzzy set, the elements are assigned a membership grade in the range [0,1]. These membership grades can have different interpretations [3]: a veristic interpretation implies that all the elements belong to some extent to the set, with the membership grade indicating the extent; whereas a possibilistic interpretation implies there is doubt on which elements belong, now the membership grade is expressing the possibility that an element belongs to the set. Last, it is also possible for the membership grades to represent degrees of truth. In [3] 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 and its membership function $\mu_{\tilde{A}}$ is given in (1)

$$\begin{aligned} \tilde{A} &= \{(p, \mu_{\tilde{A}}(x)) \mid x \in U\} \\ \mu_{\tilde{A}} &: U \to [0,1] \\ & x \mapsto \mu_{\tilde{A}}(x) \end{aligned} \qquad (1)$$

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 these criteria is a t-norm, respectively t-conorm.

T-norm	T-conorm
$T(x,y) = T(y,x)$	$S(x,y) = S(y,x)$
$T(a,b) \le T(c,d)$ if $a \le c$ and $b \le d$	$S(a,b) \le S(c,d)$ if $a \le c$ and $b \le d$
$T(a,T(b,c)) = T(T(a,b),c)$	$S(a,S(b,c)) = S(S(a,b),c)$
$T(a,1) = a$	$S(a,0) = a$

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, ...) [4].

Fuzzy sets can be defined over any domain, but of particular interest here are fuzzy sets over the numerical domain, called fuzzy numbers [5]: 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). Using Zadeh’s extension principle [1], 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*. 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 is this choice based.

2.2. Fuzzy rule base

In the fuzzy inference system, a rulebase using fuzzy premises and conclusions are used. The rulebase is comprised a set of rules that are of the form “*if x is A, then y is B*”. 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 can be a crisp value. In the rule, *A* and *B* are labels, such as “high” or “low”, also represented by fuzzy sets as described above. It is also possible to combine premises using logical operators (and, or, xor) to yield more complex rules. The “is” in the premise of the rules is a fuzzy match, implying that a value can (and most likely will) match multiple premises: a value 80 can match both “high” and “very high” albeit to a different extent. For any input (fuzzy or crisp), the process of matching the value will yield a fuzzy value indicating how well the input matches. The “is” in the conclusion is a basic assignment. 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*”.

2.3. Interpreting the output

Typical is that all the rules are evaluated and 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. As multiple rules can match, *y* can be assigned multiple values by different rules: all these values are aggregated using a fuzzy aggregator to one single fuzzy value. For each rule, the extent to which the premise matches impacts the function that is assigned to *y*. 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.

2.4. Example

Consider the simple example of a fan controller, with 3 temperature distinctions (low, normal and high). The fuzzy sets used to indicate these distinctions are shown on Figure 1. Similarly shaped fuzzy sets are used to indicate a low, normal or high fan speed. On Figure 2, the rulebase used to link the temperatures with the a speed for the fan is shown; with only a single input, the rulebase is very straightforward.

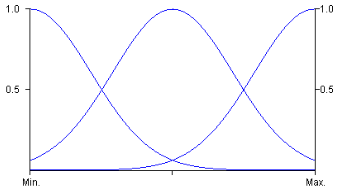


Figure 1. The fuzzy sets used to represent low, normal and high.

IF temp=low THEN fanspeed=low
IF temp=normal THEN fanspeed=normal
IF temp=high THEN fanspeed=high

Figure 2. The rulebase used for the example of the temperature and fan.

If the given input temperature matches one membership function, the outputted value of the y is exactly the function that matches. For temperatures that match multiple rules, the value of y is calculated from the output values of all the matching rules, as illustrated on figures 3 and 4.

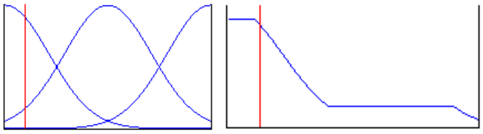


Figure 3. An input of 10 (left) and the resulting output (right) of 13 after defuzzification.

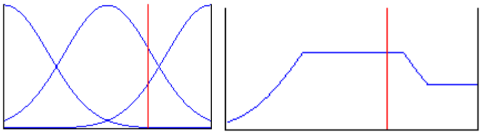


Figure 4. An input 70 (left) and the resulting output (right) of 64 after defuzzification.

On figures 3 and 4 it can be seen that the output is a fuzzy set, which needs to be defuzzified. There are several methods to define the defuzzification, and choosing a different method will lead to different – but very similar – results.

2. Application of the inference system

2.1. Conceptual example

2.1.1. Description

To illustrate the workings of the fuzzy inference system for the map overlay problem; first a simple conceptual example will be considered. The example consists of a grid comprised of two square grid tiles that holds emission data (em$_1$ and em$_2$) and a grid built up of three grid squares that holds covariate data (cov$_1$, cov$_2$ and cov$_3$); illustrated on Figure 5. Both grids cover the same area, so the different tiles don't line up properly; cov$_2$ is split into cov$_{2a}$ and cov$_{2b}$. While all cov$_i$ and em$_i$ are known, the question is how the emission values can be distributed over the grid with covariates. This problem is equivalent to correctly distributing the covariate values of a tile over its different portions: knowing how the cov$_2$ tile should be split is sufficient to derive an appropriate distribution of the related emission. In this simple example, the calculation can be done very easily; but this example will it allows us to derive the rules and verify results.

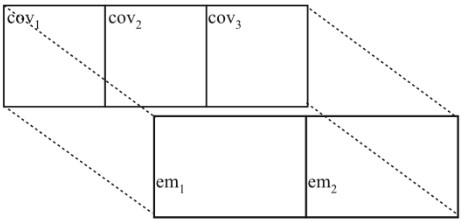


Figure 5. The grids used in the conceptual example.

2.1.2. Deriving the rules

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 covariates and emissions) are in the range 0-100. The first 5 rows show the known data; the rows cov$_{2a}$ and cov$_{2b}$ show how cov$_2$ should be distributed based on the known data.

Table 1. Examples for the conceptual dataset

em_1	100		100		0		100	
em_2	100		0		100		100	
cov_1	100		0		100		0	
cov_2	100		100		100		100	
cov_3	100		100		0		0	
cov_{2a}	50	normal	100	high	0	low	50	normal
cov_{2b}	50	normal	0	low	100	high	50	normal

In the rulebase, values are compared against predefined fuzzy sets, not against each other. To derive the rules, first assume that the covariates are equal: cov_1= cov_3. If the emission em_1=em_2, then it is obvious that cov_2 should be evenly split over both cov_{2a} and cov_{2b}. If em_1<em_2, it implies that cov_2 contributes more to em_2 than to em_1; as a result cov_{2b}>cov_{2a}. To make a rule that represents this case, we need to define the rule as:

```
IF em1=small AND em2=big AND cov1=A AND cov3=A THEN cov2a=small
```

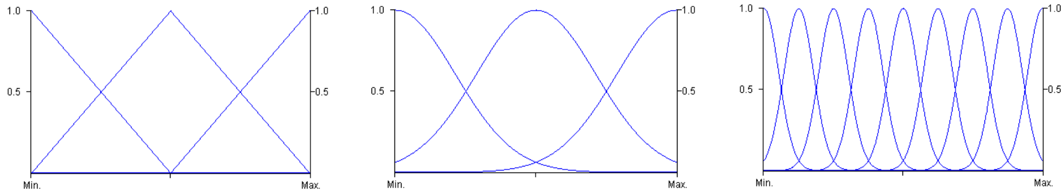
for every value of A (big, small, ...). The output value clearly depends on the difference between em_1 and em_2: the greater this difference is (em_1=very small and em_2=very big), the smaller the value of cov_{2a} should be. This yields a number of additional rules. An analogue reasoning holds when em_1>em_2.

Next, assume the emissions are equal: em_1=em_2. If cov_1<cov_3, then it implies that, as emissions are equal, cov_2 contributes more to em_1 than to em_2; so cov_{2a}>cov_{2b}; the greater the difference between cov_1 and cov_3, the more this should be reflected in the output. Consequently, we obtain the rule:

```
IF cov1=small AND cov2=big AND em1=A AND em2=A AND THEN cov2a=small
```

This is again for every value of A, and again the greater the difference between cov1 and cov_2; the more cov_{2a} should differ from cov_{2b}. A similar reasoning holds when cov_1>cov_3.

In general, neither the emissions nor the covariates will be equal. This implies that rules for those cases must be defined as well. In the current example, we considered the impact of changes to either emissions and covariates to be similar. To define the rules, we considered three predefined fuzzy sets for the emissions (representations for low, normal and high), three possible values for the covariates and nine possible values for the outputted percentage; all the fuzzy sets are shown on Figure 6. The fuzzy sets for the emissions were chosen as triangular fuzzy sets, whereas the sets for covariates and percentages are bell-shaped.



(a) (b) (c)

Figure 6. The fuzzy sets used to represent low, normal and high emissions (a); low, normal and high covariates (b) and the fuzzy sets used to determine the outputted percentages (c). For each variable, every function has a name mf_i, starting from mf_0 for the leftmost function.

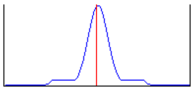
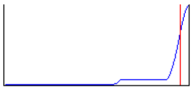
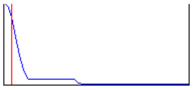
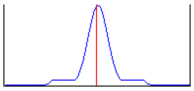
Below are some examples of the rules are shown using the fuzzy sets – the whole rulebase consists of 80 rules:

```
if (em_a == mf0 & em_b == mf0 & cov_a == mf0& cov_b == mf0 ) -> cov_percentage = mf4;
if (em_a == mf1 & em_b == mf0 & cov_a == mf0& cov_b == mf0 ) -> cov_percentage = mf5;
if (em_a == mf2 & em_b == mf0 & cov_a == mf0& cov_b == mf0 ) -> cov_percentage = mf6;
if (em_a == mf0 & em_b == mf1 & cov_a == mf0& cov_b == mf0 ) -> cov_percentage = mf3;
if (em_a == mf1 & em_b == mf1 & cov_a == mf0& cov_b == mf0 ) -> cov_percentage = mf4;
if (em_a == mf2 & em_b == mf1 & cov_a == mf0& cov_b == mf0 ) -> cov_percentage = mf5;
if (em_a == mf0 & em_b == mf2 & cov_a == mf0& cov_b == mf0 ) -> cov_percentage = mf2;
if (em_a == mf1 & em_b == mf2 & cov_a == mf0& cov_b == mf0 ) -> cov_percentage = mf3;
```

2.1.3. Examples

Using the above rulebase, we can verify some of the examples. The outputted number represents which percentage of the covariate of cov$_2$ is said to relate to cov$_{2a}$.

Table 2. Verification of the example dataset

em$_1$	100	100	0	100
em$_2$	100	0	100	100
cov$_1$	100	0	100	0
cov$_2$	100	100	100	100
cov$_3$	100	100	0	0
desired cov$_{2a}$	50	100	0	50
fuzzy result				
defuzzified	50	95.78	4.22	50

2.1.4. Remarks

Due to the fact that all the conditions are fuzzy, some results appear less optimal than we could envision them; this is mainly the case in the extreme values. Simply adding rules for the cases where one of the emissions or covariates that play a part in determining the portion is equal to 0 will not really help, as this does not prevent the other rules from matching. For a more optimal performance, testing for zero values and then applying a more customized rulebase could yield better results for those situations. For values other than these extreme cases, the outputs are nicely in between. For contradictory inputs (e.g. high emission but low covariate on one side), the results may appear a bit awkward, but this is a result of the inconsistent input.

2.2. Advanced example

2.2.1. Description

The simple example served as a means of explaining the concept. A more complicated example will be considered now. The previous example is scaled up somewhat: we now consider a 2x2 grid representing emissions and perfectly overlapping 3x3 grid containing covariates, as shown in figure 7. In this example, there basically are 3 different cases to be considered: covariate squares covered by one emission square (cov$_{11}$,cov$_{13}$,cov$_{31}$,cov$_{33}$), squares covered by two emission squares (cov$_{12}$, cov$_{21}$, cov$_{23}$,cov$_{32}$) and squares covered by

4 emission squares (cov_{22}). As the circumstances are quite different, each of these three cases will require a different approach.

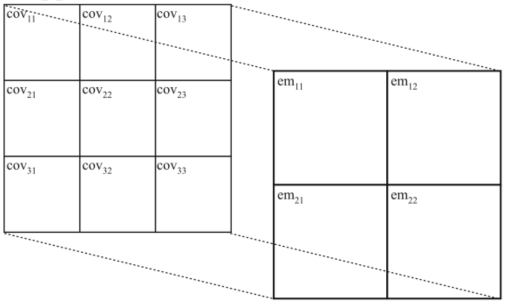


Figure 7. The grids used in the advanced example.

2.2.2. Deriving the rule bases

The concept is similar as before: the emission and covariates for known and related squares is used in the premise of the rulebase. Due to the larger nature of the example, it is impossible to consider all the possible combinations of emissions and covariates like before (this would yield 3^12 cases). Various options exist to limit the number of rules. As a simple approach, we opted to consider the ratios between emissions and covariates. As a reference to determine which ratios are high and which are low the following value is used.

$$R = \frac{\sum\limits_{i=0,1,j=0,1} \mathrm{em}_{ij}}{\sum\limits_{k=0,1,2,l=0,1,2} \mathrm{cov}_{kl}}$$

Values greater than this ratio are considered to be high, values lower considered to be low.

For the first cases, the completely covered tiles cov_{11}, cov_{13}, cov_{31} and cov_{33}, there is no need for the fuzzy inference system, as the covariate is known and needs not to be split.

For the second case, the tiles covered by 2 emission squares (cov_{12},cov_{21}, cov_{23}, cov_{32}) we need to determine which portions are relevant; we will use cov_{12} as the example (the other three are similar); and determine the value for the portion of cov_{12} covered by em_{11}. Values for the relevant ratios are needed; the neighbouring tiles that are completely covered by emission tiles are considered to determine the ratios; we will consider 2 ratios for cov_{12}. The first ratio R_1 will be defined such that it has a proportional relation to cov_{12a}, whereas R_2 will be defined to have an inverse proportional relation to cov_{12a}. As possible definitions for R_1, we have:

$$R_1 = \frac{em_{11} + em_{21}}{\mathrm{cov}_{11} + \mathrm{cov}_{21} + \mathrm{cov}_{31}} \text{ or } R_1 = \frac{em_{11} + em_{21} + em_{22}}{\mathrm{cov}_{11} + \mathrm{cov}_{21} + \mathrm{cov}_{31} + \mathrm{cov}_{32} + \mathrm{cov}_{33}}$$

The choices for R_2 are similar

$$R_2 = \frac{em_{12} + em_{22}}{\mathrm{cov}_{13} + \mathrm{cov}_{23} + \mathrm{cov}_{33}} \text{ or } R_2 = \frac{em_{12} + em_{22} + em_{21}}{\mathrm{cov}_{13} + \mathrm{cov}_{23} + \mathrm{cov}_{33} + \mathrm{cov}_{32} + \mathrm{cov}_{31}}$$

Initial tests have shown that using either definition does not make for a big difference in the end result. Note that in the above definitions only make use of the cov_{ij} that are fully covered by the emission squares considered. It is possible to also include the cov_{ij} that are partly covered by the considered emission squares definitions for R_1 could use the partially covered cov_{ij} as well, yielding

$$R_1 = \frac{em_{11} + em_{21}}{\mathrm{cov}_{11} + \mathrm{cov}_{21} + \mathrm{cov}_{31} + \mathrm{cov}_{22}} \text{ or}$$

$$R_1 = \frac{em_{11} + em_{21} + em_{22}}{\mathrm{cov}_{11} + \mathrm{cov}_{21} + \mathrm{cov}_{31} + \mathrm{cov}_{32} + \mathrm{cov}_{33} + \mathrm{cov}_{22} + \mathrm{cov}_{23}}$$

There could be similar alternative definitions for R_2, but this change most likely of little impact in the end result and would complicated things too much for a proof of concept. In the example, we will therefore consider the initial definitions.

The approach for the third situation, determining how cov_{22} should be split, is quite similar, but now different definitions for R_1 and R_2 are needed. To determine the portion of cov_{22} for the part covered by em_{11}, the following formulas will be used:

$$R_1 = \frac{em_{11}}{\mathrm{cov}_{11}} \text{ or } R_1 = \frac{em_{11}}{\mathrm{cov}_{11} + \mathrm{cov}_{12} + \mathrm{cov}_{21}}$$

The choice for R2 is similar

$$R_2 = \frac{em_{12} + em_{21} + em_{22}}{\mathrm{cov}_{13} + \mathrm{cov}_{23} + \mathrm{cov}_{33} + \mathrm{cov}_{31} + \mathrm{cov}_{32}} \text{ or}$$

$$R_2 = \frac{em_{12} + em_{21} + em_{22}}{\mathrm{cov}_{13} + \mathrm{cov}_{23} + \mathrm{cov}_{33} + \mathrm{cov}_{31} + \mathrm{cov}_{32} + \mathrm{cov}_{12} + \mathrm{cov}_{21}}$$

Using the ratios is bound to provide for less accurate results, so to compensate for this, more values have been chosen for both the relations and the covariates: we now consider 5 possible reference values for the relations, and 9 possible values for the outputted percentages, with a similar naming scheme as before. As in the conceptual example, a number of typical, predictable cases with desired results is used to derive the rulebase. For the determination of cov_{12a}, some cases are listed in the table below.

em_{11}	100	0	100	100
em_{12}	0	100	100	0
em_{21}	0	0	0	100
em_{22}	0	0	0	0
cov_{11}	100	0	100	100
cov_{12}	100	100	100	100
cov_{13}	0	0	100	0
cov_{21}	0	0	0	100
cov_{22}	0	0	0	0
cov_{23}	0	0	0	0
cov_{31}	0	0	0	0
cov_{32}	0	0	0	0
cov_{33}	0	0	0	0
R				
R_1	1	inf	1	1
R_2	inf	inf	1	inf
cov_{12a}	100	0	50	100

Based on this table, an appropriate rulebase similar to the one for the basic example is derived. Below, some of the 25 rules are listed.

```
if(R1 == mf0 & R2 == mf0) -> cov12a = mf4;
if(R1 == mf0 & R2 == mf1) -> cov12a = mf3;
if(R1 == mf0 & R2 == mf2) -> cov12a = mf2;
if(R1 == mf0 & R2 == mf3) -> cov12a = mf1;
```

The same rulebase can be used to determine the portion of cov_{22}, but of course using the appropriate definitions for the inputted relations R_1 and R_2.

2.2.3. Results and remarks

The rulebase exhibits the expected behaviour: the portion of the covariate is estimated correctly; the four example cases listed above yield results similar to the simple example. The examples are more difficult to verify though, as changing the values of the emissions and covariates for the different cases has the side effect of changing the reference ratio R. This in turn impacts the fuzzy sets used to describe high, low and so on. So far, the rulebase has been tested with simple example, but further verification is needed. As before, some very extreme cases (e.g. covariates that are 0) can yield less than optimum results, but such cases could be detected and considered separately beforehand.

3. Optimizations

3.1. Inputs

From the two examples, it obvious that the use of the ratio decreases the accuracy. Using the actual values however would yield a rulebase of unmanageable size. It may however be possible to find better groups to use (e.g. summation of emissions and summation of relevant covariates, or multiple ratios) or devise a different rulebase altogether, and obtain a better result while still keeping a relatively small rulebase.

3.2. Rulebase

In the current models, very intuitive and simple rule were used. These rules should just be seen as a first step in a proof of concept. This allows it to work for many cases but still may cause it to be less successful in other cases. The use of additional technologies (e.g. neural networks) is one approach that could allow the rulebase to be determined automatically based on a much large number of cases, rather than constructed from some intuitive results. Especially in combination with the above optimization, this should yield better results.

3.3. Use more available information

Currently, some information is not used: some covariate tiles that partly covered by an emission tile are not used. The main reason for this is that the whole point is trying to determine how to split them, but of course this may be too much of a simplification for general cases. The fuzzy inference system however allows for a more fuzzy input, which makes it possible for us to derive a representative fuzzy value for these tiles; a partially covered covariate tile could be counted as contributing it surface area (as an approximation). At present, it is not clear yet how this fuzzy value should be determined, but it will be important: the risk is that introduction more fuzzy data at the inputs could make the output value too fuzzy to be truly useful.

4. Conclusion

In this contribution, we presented a novel approach to consider the map overlay problem. To determine how data should be distributed between ill aligned grids, a fuzzy inference system is used. The methodology is still in quite early development, but is showing promising results. Future work first concerns employing a better methodology to determine and refine the rulebase and the input, and then scaling up the methodology to larger and more complex examples. Lastly, realistic examples need to be considered to verify the results in more real world situations.

References

[1] Gotway C.A., Young L.J. (2002): Combining incompatible spatial data; Journal of the American Statistical Association, June 2002, Vol. 97 No. 458, pp. 632-648.

[2] Zadeh L.A. (1965): Fuzzy Sets; Information and Control, 1 3 (1965); pp. 338–353.

[3] Dubois D., Prade H. (1997): The three semantics of fuzzy sets; Fuzzy Sets and Systems 90, pp. 141-150.

[4] Dubois D., Prade H. (2000): Fundamentals of Fuzzy Sets. Kluwer Academic Publishers.

[5] Klir G. J., Yuan B. (1995): Fuzzy sets and fuzzy logic: Theory and applications; New Jersey: Prentice Hall.

[6] Zimmerman H-J. (1999): Practical Applications of Fuzzy Technologies; Kluwer Academic Publishers.