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 4: A fuzzy rulebase approach to remap gridded spatial data: initial observations

Jörg Verstraete

Systems Research Institute - Department of Computer Modelling
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
jorg.verstraete@ibspan.waw.pl
http://ibspan.waw.pl

Abstract. In many fields of research where different gridded spatial data needs to be processed, the grids do not align properly. This can be for a multitude of reasons, and it complicates drawing conclusions and further processing the data; it requires one grid to be transformed to match the other grid. In this article, we present the first results of a completely new approach to transforming data that are represented in one grid, to have it match a given target grid. The approach uses techniques from artificial intelligence and simulates an intelligent reasoning on how the grid can be transformed, using additionally available information to estimate the underlying distribution. The article describes the algorithm, and results on artificial datasets are discussed.

1 Introduction

1.1 Problem description

Numerical data that are spatially correlated are often represented in a gridded format. This means that the map over which the numerical data holds, is divided using a raster. Each cell of the raster (or grid) is then assigned a value that is deemed representative for this area. As such, the real world spatial distribution of the modelled value is approximated using a discrete model. Usually, a regular grid with rectangular or square cells is used. Data are often supplied from different sources, and different data are acquired using different technologies. As such, the data are often represented in *incompatible* grids: these are grids that have a different orientation, or different size of grid cells. They are called incompatible, as it is not possible to directly map data from a cell in one grid, to another cell in the other grid. However, this is exactly what needs to be done: scientists want to find correlations between two grids, or assess the influence of one feature onto another feature (e.g. the concentration of air pollutants to which people are exposed). One example is the health impact of airborne pollutants, such as described in [1]. A more complicated example would be judging the benefit of cycling in a city [2]: cycling is good for your health, as it is physical exercise, but cycling in a polluted environment may cause more downsides than benefits.

There is the exposure to exhaust gasses, but also the changed risk and effects of having an accident, which also needs to be taken into account. Such studies require pollution data, traffic information, accident statistics, traffic patterns and many more. All this information is usually not represented in the same format, and combining the data properly is an issue.

Consider for instance a pollutant that is present over a large area, most likely in different concentrations at different places. The exact distribution might not be fully known (e.g. due to a limited number of measuring points) and is provided as a regular grid with grid cells of e.g. 500m x 500m. Similarly, the population density can also be provided in a gridded format, but its grid cells can have a different size, e.g. 100m x 100m, and even be rotated. Determining which people are exposed to which concentration is a complicated problem, and requires transforming one grid onto the other one. This is illustrated on figure 1: a 4x4 grid has to be remapped onto a 9x9 grid that is slightly angled. If it would be known that the data is related to the black line, the distribution in the 9x9 grid can be better suited, as shown by shaded squares in the examples (a) and (b). Current methods often result in transformations in which the data is more spread out, and moves away from the likely underlying distribution. To overcome this, we present a method that incorporates additional information in order to perform a better transformation of the data.

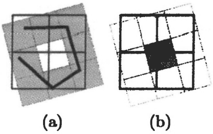

Fig. 1. Example of an input grid (2x2, in bold) that needs to be remapped onto a target grid (3x3, dotted line). Different additional data are represented by the thick lines in (a) and (b).

1.2 Current solution methods

Current solution methods all work on the same principle: the grid that needs to be transformed (this will be called the *input grid*) is analysed and a distribution of the underlying data is assumed. Using this assumed underlying distribution, the data are then remapped to match the grid it will need to be combined with (this will be the *target grid*). A short summary of the most common methods is supplied; for a more detailed overview on different approaches for the map overlay problem, we refer to [3].

The most commonly used is *areal weighting*, [4]. In this approach, the amount of overlap between a gridcell of the input grid and a gridcell of the target grid determines the portion of the associated value of the input gridcell that will be remapped onto the target. Each target gridcell thus gets associated a weighted sum, where the weights are determined by the amount of overlap with the overlapping input gridcells. This approach assumes that the data in each cell of the input grid are spread uniformly. This assumption however is not always true: in the case of air pollution, the concentration of some pollutants could be linked to linear sources e.g. traffic on roads or could be caused by point sources, which implies that the concentration should be more focussed around the source (taking into account dispersion of the pollutant using existing dispersion models). In figure 1, this means that the data of the 4 input grid cells would be spread out over the 9 target grid cells, not considering the source indicated by the thick line.

A more refined approach to this is *areal smoothing*. In this approach, the data modelled in the input grid is approximated by interpreting the data as a third dimension, and fitting a smooth surface over it. The assumption here is that the data modelled by the input grid are showing a smooth distribution over the whole region of interest. The smooth 3D surface is then resampled using the target raster. This sampling results in the values that will be associated with the cells. While allowing for a smooth transition, the method has the same disadvantage as areal weighting, in that it cannot cope well with local effects such as point or line sources.

2 Rulebase approach

2.1 A different look at the problem

The main issue with the problem is that the underlying distribution is not known: the current methods approach the problem by (implicitly) assuming a distribution. Additional knowledge might however be present to help determine a better distribution. An example where one dataset can be improved is when different datasets are fused. In [5], the authors combine different datasets in order to obtain a higher quality dataset. The methodology however is applied on vectorial data that is tagged (e.g. a region tagged as forest, a region tagged as agricultural land, etc). After deriving a common ontology, and after combining the different definitions for regions on the maps, the authors derive a new map that contains the combined information of both.

Generally, when there is a grid representing data, there might be other knowledge that are known to influence the distribution. In the ongoing example of the air pollutant, the type of pollutant and its source can provide information on this. If the particular chemical or particle originates from car exhausts, then the distribution should more or less match the road network (after correction for dispersion). Different pollutants might as such have a different underlying distribution. Such knowledge, makes it possible to make good judgements on the underlying distribution, as shown in [6]. For every grid cell in the target grid, the additional knowledge can be considered. This can be by taking amount over overlap with features of the additional knowledge, the distance to specific items, etc. In [7], a detailed description on how an expert would reason about the redistribution using additional data is presented.

The additional knowledge should only be used to *steer* the distribution, but it should not be followed too strongly: if the additional knowledge is from a different time (e.g. pollution data from 2010, traffic information from 2009), the correlation is weaker. Following that data too strongly might not even be possible and thus would either yield no solution, or a solution that obscures real data. The additional data might also not be the only explanation for the distribution, other sources might be present but unknown. This again is an argument for not too strictly adhering to this information.

2.2 Emulating the intelligent reasoning

A fuzzy inference system is a system that uses fuzzy sets to represent data and evaluates predicates using simple rules and fuzzy matching [8]. Fuzzy sets are a way of representing uncertain or imprecise information by means of a membership function ([9], [10]). The membership function indicates the possibility or membership of each value. Given an adequate domain, such membership functions can be used to represent e.g. linguistic terms such as *low*: on a domain $[0,100]$ all values below 10 can the be low (with possibility 1), values above 20 can be considered as *not low* (represented by possibility 0), and values between 10 and 20 have a linearly decreasing membership from 1 to 0. The fuzzy inference system has multiple rules of the form:

`IF x is <linguistic term> THEN y is <linguistic term>`

Here, `x` is a numerical input variable, `y` is the output value, and `<linguistic term>` is a fuzzy set representation for e.g. high, low or other other possible value descriptions. There can be multiple input values, combined using logical operators *and* and *or*. The input variable is matched against the linguistic term, which results in a value in $[0,1]$ that indicates how well the value matches the term. Based on this, `y` is assigned a linguistic term in its domain. The term is represented by a fuzzy set. There are multiple rules in the rulebase, and `x` can match multiple rules at the same time, resulting in multiple fuzzy sets for `y`. All these results for `y` are aggregated to a single fuzzy set, which is then subsequently defuzzified to yield the crisp result. Several algorithms for defuzzification exist, but for now the most common *center of gravity* will be used.

In [7], we presented how an inference system can be applied to emulate the intelligent reasoning. Key to achieving this is defining the rulebase and the parameters in the rulebase. In order to guarantee that the new distribution still resembles the input distribution, the redistribution of the data happens locally, within a single input cell. The target grid is specified completely independent from the input grid, so first a new grid is computed, the *segment grid*. This grid is made up of all the intersections between input and output cells. Each cell in this grid (for the remainder of this article called *segment*) will only overlap with a single cell from the input grid, and with a single cell from the output grid. Every input cell is completely and exactly covered by a number of segments, as is every output cell. In the algorithm, the segment grid will be used as the new target grid. The problem then becomes a problem of redistributing the data in an input cell over the contained segments. Subsequently, the segments can be combined differently to form output cells. To facilitate implementation, the additional knowledge is also represented as gridded data. Even if the original knowledge is in a different format (e.g. a road network represented by lines), it is a straight forward operation to convert this to a grid with a small cell size.

2.3 Parameters and range

In order to make the inference system, it is necessary to define parameters. These are values that are considered to provide some correlation with an output: proportional (a high value of the parameter coincides with a high value of the ideal value), or inverse proportional (a higher value of the parameter coincides with a lower value of the ideal value). In [11], several candidates for parameters were proposed. Here, the considered parameters are:

- amount of the auxiliary cell covered by the segment
- amount of the input cell covered by the segment
- amount of the interior of the auxiliary cell covered by the interior of the segment

These parameters were chosen after running several experiments, as they provided the best overall results. Consider the first parameter: "amount of the auxiliary cell covered by the segment". It is intuitive to state that the more of the auxiliary cell is covered by this segment, the higher the value of this segment should be: higher auxiliary value should yield a higher output value. In the rulebase this could be called **aux_overlap**, the value would be used in a rule of the form:

```
IF aux_overlap is low THEN output is low
IF aux_overlap is medium THEN output is medium
IF aux_overlap is high THEN output is high
```

The linguistic terms low, medium and high for **aux_overlap** need to be defined, which means finding adequate limits for the domain of the **aux_overlap** value. When the limits of the domain for the parameter (e.g. **aux_overlap**) are known, a number of equally spaced and shaped triangular fuzzy sets are defined over this domain to define the linguistic terms. The number of triangular fuzzy sets is chosen for each parameter. The more fuzzy sets are defined on the domain, the more rules the rulebase will have; this effectively poses a practical limit. More fuzzy sets should yield more possibilities of distinguishing different values. The main problem now is determining the domain. In our interpretation, the domain is defined by the possible values this particular parameter can have for this segment, thus it varies with each segment. For the relation between segments and the auxiliary grid, there are several possibilities. In the simplest case, the segment covers part of an auxiliary cell. The total value of this auxiliary cell can therefore be considered to be in this segment (e.g. in case of a point source that is the source of the entire value), partly in this segment, or not at all in this segment (e.g., if the value is due to sources outside of this segment - which is possible as there are other segments overlapping the same cell). The first case results in the maximum possible value. The last case results in the minimum possible value, 0 unless one or more auxiliary cells are fully contained inside the segment, the minimum possible value is then total value of those contained cells. The weighted value is considered as the value of the parameter that is verified, and thus is passed as parameter **x**. The calculation for the range of other parameters is done similarly.

The range for the value of an output cell is initially unknown, but it is limited by the total of its containing segments. For each segment, the output range is from 0 to the value of the overlapping input cell - due to the definition of the segments, there is exactly one. The exact value is obtained using the fuzzy inference system, resulting in a fuzzy set that is then defuzzified. However, the values of all segments that form an input cell should sum up to the value of that input cell. As the fuzzy result for each segment is defuzzified independently, there is no way to guarantee this. Currently, the defuzzified output is considered as a proportion of the total value of all segment: the real range does not matter, so for now the output range for each segment is $[0, 100]$, where 100 is an arbitrarily chosen value. Once all the segment values are calculated and defuzzified, the obtained value is interpreted as a relative amount of the total of all values for segments that overlap this input cell.

2.4 Rulebase construction

The construction of the rulebase at present is fairly rudimentary: after rules that evaluate the first parameter, the addition of each new parameter multiplies the number of rules by the number of linguistic terms for that parameter. It makes every possible combination of each linguistic term for this parameter and the existing rulebase. In the current examples, three parameters, each represented to ten linguistic terms, result in a rulebase that has 10^3 rules. The range of the output value is expanded with each added parameter: the more parameters say it should be a large value, the larger the value will be. Afterwards, the output range will be rescaled to match the true range. This is a simple way of creating the rulebase, but it results in a very big rulebase in which many rules may never be matched: contradictions between parameters are present.

3 Experiments

To test the methodology, different datasets were generated: a geometric test pattern was sampled onto an 12x12 grid (figure 2a) to result the input grid. In the first three test cases, the grid has to be remapped onto a 25x25 grid; the optimal solution (obtained by sampling the geometry onto the grid) is shown on figure 2b, the solution using areal weighting is shown on figure 2c. The fourth test case requires the remapping onto a 25x25 grid that is at a 20° angle, the

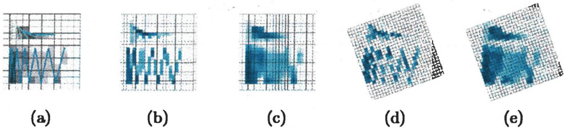

Fig. 2. (a) generated input data with grid, (b) ideal solution for target 1, (c) areal weighting for target 1, (d) ideal solution for target 2, (e) areal weighting solution for target 2. Darker shades represent higher associated values; but the scale between different grids does not match. For each grid, black indicates the highest occurring colour in that grid; the lighter the colour, the lower the associated value.

ideal solution and areal weighting solution are shown on respectively figure 2d and figure 2e.

All samples were run using the same three chosen parameters from the previous section; the rulebase system was generated in the same way for all tests, and used ten linguistic variables defined over the domains of each parameter.

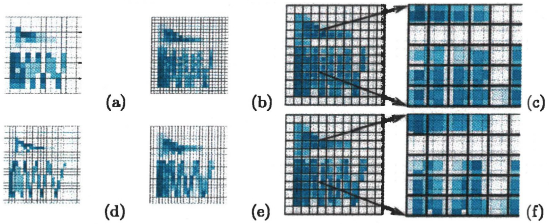

Fig. 3. Case 1: low resolution auxiliary data: (a) auxiliary data, (b) result, (c) detail of the remapping of the input data and case 2: High resolution auxiliary data: (d) auxiliary data, (e) result, (f) detail of the remapping of the input data.

The developed methodology uses auxiliary data that has to be provided by the user. Experiments were run with different data as auxiliary data, but the auxiliary data was also presented on a grid which was sampled from the same geometries as the input data: this yields *perfect* data, which should provide the best results and allows for the system to be tuned and verified.

In the first test case, 15 × 15 auxiliary grid with the same orientation (figure 3a) is used. The result (figure 3b) clearly reveals more detail than areal weighting (figure 2c). The second test case uses a 27 × 27 auxiliary grid (figure 3d), and the result shows even more detail (figure 3e). As input and target are the same, it should be compared against the same areal weighting result. The redistribution of the data in the input cells over the segments are shown on figure 3c and figure 3f for both cases: the bold lines show the input grid, the dotted line the output grid. The segment grid is the irregular grid defined by all these lines. The center part of the segment grid is enlarged for better visibility. On the segment grids, it is clear to see how the value of each input cell is redistributed over its segments. The benefits of the higher resolution auxiliary data are even more visible on this segment grid. Particularly in the second column of the input, the redistribution differs as a result of the different auxiliary data, and the higher values are shifted to the left of those cells. The third test case uses the same target grid, but now employs an auxiliary grid 15 × 15 angled 10°. The fourth test case illustrates the effects if the 25 × 25 target grid is angled 20°. Particularly the distribution of the data inside the cells over the segments in interesting, figures 4c and figure 4f.

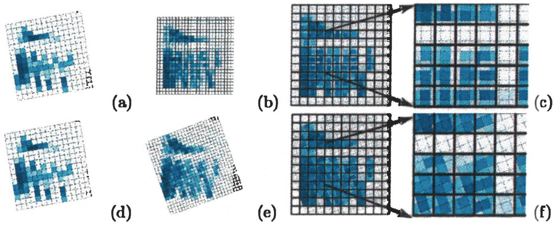

Fig. 4. Case 3: low resolution rotated auxiliary data: (a) auxiliary data, (b) result, (c) detail of the remapping of the input data and Case 4: low resolution rotated auxiliary data and rotated target: (a) auxiliary data, (b) result, (c) detail of the remapping of the input data.

To derive a more quantified comparison, consider the absolute difference of the value of a cell in the optimal grid (figure 2b for the first three cases, figure 2d for the fourth case) and the calculated value for the same cell. In table 1, the average and maximum of these differences are shown for both the presented method and for areal weighting. The values for average weighting are the same for the first three cases, as the input and targets are the same. For the results of the presented method, all the values of the second case are lower than those of the first case. This means that the second case has a better result, which is also visible on the figures (figure 3b vs. figure 3e). For these cases, the presented method has a lower average difference than areal weighting, but it has a higher maximum average. In simple terms, this means that there are less errors, but larger errors occur. This is consistent with the fact that our methods concentrates the data more, whereas areal weighting tends to smear out the data more over multiple cells: where the line patterns using areal weighting is just visible as a blur, the presented method is able to distinguish more of the pattern. In the case

	average difference		maximum difference	
	presented	areal weighting	presented	areal weighting
case 1	3.70	3.97	41.62	33.16
case 2	3.11	3.97	39.74	33.16
case 3	3.54	3.97	32.86	33.16
case 4	7.32	7.55	81.00	82.25

Table 1. Properties of the results of the 4 examples.

3 and 4, a low resolution auxiliary grid was used to show that this is enough to contribute. A 15 × 15 grid does not add that much information over a 12 × 12, but still enough to provide better results. Case 3 shows that the low resolution auxiliary grid at an angle performs slightly worse on average, but better on the maximal difference. In case 4, the values are much higher, as remapping to an angled grid is a more complicated issue. But the presented method still outperforms areal weighting. Compared with the areal weighting approach, the proposed methodology offers better results in remapping the data to the target grid, even when the auxiliary data has a relatively low resolution. The segment grids provide the highest resolution, but unfortunately are irregular. Particularly when input and target are at an angle, the resulting segment grid is not suitable as final representation. The conversion to the target grid is done by adding up all segments that together belong to the same grid cell in the target grid. This effectively lowers the resolution again, which is clearly visible on the figures of the segment grid. However, it results in the desired format. This final step is irreversible: it disconnects the result from the original input grid, and by adding up the values of the segments, the value of an input cell is possibly spread out over a slightly larger region.

4 Conclusion

The complexity of the presented method is linear with the number of cells in the segment grid, i.e. the number of cells in the intersection of input and output grid. Consequently, the method scales quite easily. Furthermore, the calculation of each segment can be done independently of other segments, implying they can be computed in parallel. In the above examples, the parameters were manually chosen by us from a large set of parameters ([11]), based on empirical studies on many data. Automatically determining the best possible parameters for a given dataset would improve the applicability. As can be seen on the segmented grids of all examples, but more-so on figure 4, all calculations are constrained within the cells of the input grid.The method tries to localize point sources or line sources at a local level. Mapping the data from the segments to the target grid has the result that data of a segment is spread out over a larger area. As such, it may also give the impression that data are moving out of the original input cells, particularly as the resulting grid is later most likely interpreted as having a uniform distribution within the grid cells. The same applies however to other methods, but as the intermediate results show higher accuracy, perhaps a different aggregation can be considered. In the presented approach, each cell from the input grid is divided in a number of segments, a possibility distribution for the value of each segment is determined. The value of all segments overlapping an input cell should sum up to the value of the input cell; to achieve this, the defuzzified results were interpreted as relative portions, which required an additional rescaling. The results can be improved by performing a more appropriate defuzzification, and avoiding the rescaling.

In this article, we presented the first experimental results of a novel way to transform gridded data. Unlike current methods, the approach uses additionally known information to estimate an underlying distribution. The presented method uses a fuzzy inference system in order to determine the values of the grid cells in the target. The results are promising, and further research in both refining the system and testing it with real world data are foreseen.

Acknowledgement The study was conducted within the 7FP Marie Curie Actions IRSES project No. 247645, acronym GESAPU.

References

1. M. Tainio, M. Sofiev, M. Hujo, J.T. Tuomisto, M. Loh, M.J. Jantunen, Karppinen A., L. Kangas, N. Karvosenoja, K. Kupiainen, P. Porvari, and J. Kukkonen. Evaluation of the european population intake fractions for european and finnish anthropogenic primary fine particulate matter emissions. *Atmospheric Environment*, 43(19):3052–3059, 2009.
2. James Woodcock, Marko Tainio, James Cheshire, Oliver O'Brien, and Anna Goodman. Health effects of the london bicycle sharing system: health impact modelling study. *British Medical Journal*, 348, 2 2014.
3. Carol A. Gotway and Linda J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.
4. R. Flowerdew and M. Green. *Spatial analysis and GIS; eds. Foterhingham S. and Rogerson P.*, chapter Areal interpolation and types of data, pages 141–152. Taylor & Francis, 1994.
5. M Duckham and M. Worboys. An algebraic approach to automated information fusion. *International Journal of Geographic Information Systems*, 19:537–558, 2005.
6. Jörg Verstraete. Using a fuzzy inference system for the map overlay problem. In *3rd International Workshop on Uncertainty in Greenhouse Gas Inventories*, pages 289 – 298, 2010.
7. Jörg Verstraete. Solving the map overlay problem with a fuzzy approach. *Climatic Change*, pages 1–14, 2014.
8. Jerry M. Mendel. *Uncertain rule-based fuzzy logic systems, Introduction and new directions*. Prentice Hall, 2001.
9. Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.
10. Didier Dubois and Henry Prade. The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, 90:141–150, 1999.
11. Jörg Verstraete. Parameters to use a fuzzy rulebase approach to remap gridded spatial data. In *Proceedings of the 2013 Joint IFSA World Congress NAFIPS Annual Meeting (IFSA/NAFIPS)*, pages 1519–1524, 2013.