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 3: Automatically identifying suitable rulebase parameters in the context of solving the map overlay problem

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. Analysis of geographically related data often requires the combination of data from different sources. Data are commonly represented in grids, and unfortunately, the grids containing different data do not match properly: they can differ in cell size and/or orientation. A novel methodology was presented to allow the data of one grid to be remapped onto the other grid. The method makes use of a fuzzy inference system that performs the remapping, using additional information relating to the data distribution. Previous research has revealed that the best parameters used in the inference system are dependent on the input, and as such an automatic determination of which parameters should be used, would improve the performance. In this article, we propose a solution for this automatic detection, by first generating a training set that is related to the input and then determining what the best parameters are for this training set.

Keywords: map overlay, spatial reasoning, fuzzy inference system

1 Introduction

In geographic sciences, there often is a need to combine data coming from different sources. Combining such data poses interesting problems: the data are obtained using different technologies and commonly the format in which the data are presented differs. A common representation format for numerical data spread over a region (such as e.g. concentration of a pollutant) is a grid: a raster that divides the region of interest into a number of cells, each of which is assigned a value that is considered to be representative for the area covered by that cell. However, different data can be defined on incompatible grids, these are grids between which there is no one-to-one mapping of the cells of the grids. Consequently, it is very difficult to compare the different grids, and to draw conclusions. As there is no clear mapping of one cell to another cell, it is difficult to perform correct calculations and to draw reliable conclusions. This is for example the case when studying the exposure of humans to specific airborne pollutants, where the population data does not align properly with the pollution

data, but it also occurs in many environmental or wildlife studies. In [1], a novel methodology to pre-process gridded data to help solve this issue was proposed. This approach uses a fuzzy inference system in order to derive new output values. The inference system requires parameters that relate to the output; several candidate parameters were presented in [2]. Initial observations presented in [3] showed that the performance of the system is highly dependent on the parameters used. In [3], these parameters were chosen intuitively and manually in order to judge the performance under ideal conditions. In this article, a methodology to find ideal parameters using an automatically generated training set for a given dataset is presented and verified using experiments.

2 Problem description

2.1 Map overlay problem

Combining data that are represented on different grids implies combining the data in the cells of the grids. The first issue is that there are many ways to define a grid: the size of the grid cells can differ, or one grid can be at an angle when compared to another grid. Grids with different orientation or size of grid cells are called *incompatible*, which makes it very difficult to compare data in one grid to data on the other grid. This is further complicated by the second issue: in a grid, the grid cell is considered to be the smallest unit for which data are known. With each grid cell, a number, representative for the cell, is associated. Usually, this is an aggregated value, but the underlying spatial distribution that resulted in the value is not contained in the data. If the number for instance holds the concentration of a pollutant, there is no way of knowing how it is distributed inside the cell: there can be a single point source, or it can be uniformly distributed over the area covered by the cell. The fact that the distribution is not known makes it difficult to map cells of incompatible grids to one other; the problem is referred to in literature as the *map overlay problem*. The general approach to solve the map overlay problem, is to remap one grid onto to other grid (sometimes implicitly), in order to achieve a one-to-one mapping between the grid cells.

2.2 Current approaches

General concept To find a clear mapping between different grids, the most straightforward idea is to transform one grid onto the other grid; this means remapping the data contained in one grid, to match the other grid. This makes the grids compatible, and results in a one-to-one mapping of the grid cells of both grids. Different methods exist, an overview of current solutions is in [4] and in [5]. A short overview is listed below.

Areal weighting In areal weighting, the data represented within each cell is considered to be uniform over the cell, and independent of neighbouring cells.

The contribution of a cell of one grid to a cell in the other grid is determined by the amount of overlap: the percentage of overlap is the percentage of the modelled value that is mapped in the other grid cell.

Areal smoothing In areal smoothing, the modelled feature is considered to be smooth over the area; mathematically this is achieved by interpreting the modelled feature as a third dimension and fitting a smooth surface over the volume. Resampling this smooth surface using a different raster, results in the remapped grid. In both approaches, assumptions regarding the underlying distributions are made, but these assumptions very often have no connection to the real world situation.

Regression methods Different regression methods to approach the problem exists. In these methods, an attempt is made to establish patterns of overlap, which are then used to estimate the data. The data are often assumed to have a specific distribution (e.g. poisson). The assumption of the distribution is also here what limits the possibilities of this approach.

2.3 Additional data approach

Data fusion is a field in which different datasets are combined with the aim of providing a higher quality dataset. In [6], the authors combine datasets that contain descriptive data: regions on the map are annotated with text labels to indicate types of land cover. The definition of the regions differ, as do the text labels used, yet it is possible to combine the knowledge to yield a better data.

When considering a grid with numerical data, often other data are available that are known to be related to this data. In the origin of our research, the numerical data concerns concentration of specific airborne pollutants. Other studies have shown a correlation between the presence of these pollutants and traffic. Consequently, the distribution of the pollutant should relate to the road network and traffic density - taking into account dispersion of the pollutants, for which we have a dispersion model available. In [1], it was proposed to use this additional information: the transformation of a grid that contains data on such a pollutant might be done better when taking the road network into account. However, using additional information poses new problems. Even though a correlation between both supplied data (particular pollutant and traffic) might be known, in general, the data will be from different sources and may concern data measured at different times or with different accuracy. Furthermore, there is no guarantee that the supplied additional knowledge is the *only* explanation for the original data: there may be other sources of this particular pollutant that are not known or supplied. The additional information can therefore only be used to supply information on the underlying distribution. Following this additional information too strictly might yield no solutions or might obfuscate other sources. Despite these uncertainties and imprecisions, it is possible to perform an intelligent reasoning in order to achieve a better distribution. The intelligent

reasoning is further elaborated on in [1], the subsequent intelligent method and initial results are described in more detail in [7].

3 AI Algorithm

3.1 Prerequisites

The algorithm presented in [7] processes the data using a fuzzy inference system, in a way that mimics the intelligent reasoning. The fuzzy inference system is a concept from artificial intelligence, in which fuzzy sets are used to determine new values. Fuzzy set theory is an extension of traditional set theory which, among other things, allows for the representation of uncertain or imprecise values. This is achieved through the use of a membership function, which maps the domain onto the interval $[0, 1]$

$$\begin{aligned} \tilde{A} &= \{(x, \mu_{\tilde{A}}(x)) | x \in A\} \\ \mu_{\tilde{(A)}} : A &\rightarrow [0, 1] \\ x &\mapsto \mu_{\tilde{A}}(x) \end{aligned}$$

Higher membership grades imply higher possibilities or certainties - this depends on the interpretation given to the fuzzy set ([8]). A consequence of the ability to represent imprecise values, is that fuzzy sets also can be used to represent linguist terms such as *high* or *low*, by defining a linguistic term as a fuzzy set over the domain and associating higher membership grades to values of the domain that better match the linguistic term. If for example values range between 0 and 100, the linguistic term representing high can be represented by a fuzzy set that increases linearly from 0 for the value 80, to 1 for the value 100. The core of the fuzzy inference system is a set of rules of the form

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

Here, `x` is an input variable that it matched with the linguistic term; the first `is` is a fuzzy match that matches a numeric value with a fuzzy set (which is the representation of the linguistic term). As such, the rule is a representation for a natural language predicate, e.g. *if x is high*. Multiple parameters and linguistic terms can be combined through logical operators such as *and* and *or*, to form a more elaborate premise. The `y` is the output value that is assigned a value, which is a linguistic term also represented by a fuzzy set. The inference system has multiple rules, and typically all these rules are evaluated. As multiple rules can have a matching premise, there can be multiple values for `y`. These are aggregated using a standard fuzzy aggregation method to yields a single fuzzy set that represents the output value. This is then defuzzified to result in the crisp, numerical value returned by the system.

3.2 Translating the problem to fit a rulebase

Parameters In [7], it is explained how the given problem is translated to fit the rulebase approach. The rulebase system can be considered separately for each cell in the output grid; the output value y is the value that should be associated with the cell. To employ the rulebase, it is necessary to find parameters x that relate to the ideal output. With these parameters, it is possible to generate a rulebase that will compute an output value y. For a given output cell, one example of such a parameter would be the total value of the auxiliary cells that overlap with the output cell. In [2], different parameters were proposed, some based on overlapping cells, others based on distances. As the value of the parameter will need to be matched against linguistic terms, an adequate range for each parameter is also needed: this range allows us to defined the linguistic terms, and thus to say when a value is high or low. In [2], a number of intuitively obtained parameters were proposed and manually verified. In [3], several parameters and their ranges were considered, and simulations were run on artificial data. For each segment, every parameter and its possible range (which is specific for each segment) were calculated. The parameters were manually selected and used in the fuzzy inference system to determine the underlying distributions. The simulations showed that the performance of the system is different per case, and that different datasets benefit from a different selection of parameters. Consequently the performance of the system can be improved by finding the most appropriate parameters for a given dataset. However, even though the target grid is supplied, directly calculating new values for the cells of the target grid is problematic: a cell in the target grid can overlap with multiple input cells. This makes it difficult to determine its value, as different input cells are involved, and the portion of each input cell that should be mapped to the output cell needs to be found.

Segments Basically, the goal is to redistribute the data within each input cell, and then remap it to the target grid. This can be done by considering a new grid, obtained from the intersection between input and target grid, referred to as the *segment grid*. The segment grid is an irregular grid, which has the property that every segment belongs to exactly one cell in the input grid, and to exactly one cell in the output grid. Furthermore, every cell both in input as output grid is covered by an integer number of segments. Examples are shown on Figure 2(a) and 3(b). Consequently, it is possible to redistribute the data of an input grid cell over the segments that are covered by it. After this, the value of a grid cell in the target grid can be computed by aggregated the the different segments that it covers. For the rulebase system, the segment grid will be used as the target grid.

4 Parameters

4.1 What are the *best* parameters?

In [2], a number of parameter definitions were proposed. These range from quite intuitive values (e.g. amount of overlap between a segment and the additional data), to more elaborate ones (e.g. the distance to high values in additional data that do not overlap the overlapping input cell). Not only is it necessary for the calculated parameter to relate to the optimal output but, it is also necessary to find an appropriate range in order to assess when the parameter is high or low. Criteria for good parameters are:

1. relate to the output value
2. have a proper range

4.2 Relating to the output

The first requirement of a good parameter, is that it relates to the output value, yet in general, the output value is unknown. Suppose that output value is known, then it is possible to use the Pearson correlation (1) to determine for each parameter whether or not it relates to the ideal output, and how well it relates. Pearsons's product-moment correlation between a list of values X and a list of values Y is defined as ([9]):

$$cor(X,Y) = \sum_i \frac{(x_i - E(X))(y_i - E(Y))}{(n-1)s(X)s(Y)} \tag{1}$$

Here, $E()$ is the expected value, approximated by the mean of the list , and $s()$ the notation for the standard deviation of the list. The Pearson correlation results in a value in the range $[-1, 1]$. Positive numbers imply a proportional correlation (thus for the parameter, this implies a proportional connection: high values relate to high values) and negative numbers an inverse correlation (thus high values relate to low values). The closer the value is to 0, the weaker the correlation. The Pearson correlation therefore not only provides data how the parameter relates to the output, but also how well it relates. The problem with using the Pearson correlation is that it requires a data set in which ideal output values are known, which moves the problem of finding good relating parameters to finding a suitable training set. This will be considered in Section 5.1.

4.3 Proper range

After the first stage of determining what the good parameters are, it needs to be investigated if an appropriate range for the parameters can be defined. A parameter that shows perfect behaviour compared to the ideal output is useless without a properly defined range, as there would be no frame of reference to know if the value is high or low in the rulebase system. The range of a parameter should reflect the possible values of the parameter. The lower bound of the

range is defined as the lowest possible value that this segment can have while still maintaining a possible remapping. The upper bound is defined as the highest possible value that this segment can have while still maintaining a possible remapping. An appropriate range has the following properties:

1. it is not a degenerate interval for *all* segments
2. it is such that the parameters for different segments do not have the same relative value within the range

If the range is degenerate, i.e. lower and upper limit are equal (and thus equal to the parameter), the evaluation to high or low will yield the same result, and this parameter will not contribute to the outcome. However, the range can be degenerate for some segments: the parameter would still contribute for the segments for which it does have a valid range. If a parameter evaluates the same in the range for every segment, then this parameter does not contribute either. All the evaluations for the parameter will be the same (it will always be low, or always be high, to the same extent), for every segment. This property is more computationally intensive to verify, as it implies evaluating the parameter with the range and comparing the outcome for all segments.

4.4 Example

A simple example for a parameter is the value represented in the overlapping cells of the auxiliary grid. For a given output segment, the value of the parameter `x` will be the weighted sum of all overlapping auxiliary cells, where the percentage of overlap are the weights; this is the value that would be assigned to the segment in the case of areal weighting. The range is considered specifically for this parameter and this output segment, and will be the possible range this parameter can have. The highest possible value occurs if the value of the overlapping auxiliary cells is completely mapped onto this segment. This simulates the situation where the explanation for the value of these cells is fully located inside the selected segment, and this provides an upper limit. The lowest possible value is the inverse situation: the justification for the value of the overlapping auxiliary cells is not in the segment, in which case the lowest possible value is 0. It should be noted that if cells of the auxiliary grid are fully contained inside this segment, then the sum of their values serves as the minimum possible value (this data can never be mapped outside the segment). Other parameters and ranges are calculated in a similar way.

5 Parameter selection

5.1 Data generation

In order to select which parameters are suitable using the Pearson correlation, it is necessary to obtain data that is representative for the given problem, but also has an ideal output.

In [3], the author concluded that the parameters are highly dependent on the dataset used. Additional experiments showed that the main reason is the difference between the rasters: how are the grids oriented and what are the relative cell sizes? It was observed that - as long as the rasters are similar - the ideal parameters are usually the same[1]. As such, it is sufficient to generate a training set that has the same grid definitions as the supplied problem. It still needs values for different cells; to achieve that, geometries as shown on figure 1c were generated and values were associated with the features. The positioning and values of the features was not random, but followed specific rules to force a variety in the numerical data. The main argument for not generating the training using fully random data to prevent situations where the randomness might yield less adequate training set, and consequently to also make sure the developed system is deterministic. The pattern was chosen so that it is easily accommodated for different grid definitions; it also provides a subjective view on how good the grids are an approximation of the situation. The pattern was then sampled using the provided grids.

5.2 Calculation

In the test data, it is possible to calculate every parameter and its range, for every output segment. As the test pattern was also sampled with the segment grid, there is an ideal value for each segment. Consequently, the Pearson correlation can be calculated for each candidate parameter, comparing its values with the ideal output values. The best parameters are those with the correlation values furthest from 0.

Note that it is also possible to calculate the range on the training set, and only keep those parameters that have a good range. However, the range is specific to the value of a parameter and segment in a given situation and its computation does not need the ideal output value. There actually are several benefits to calculating the range on the original dataset rather than on the training set. First, it will allow to select those parameters that have a proper range for the current problem. Even though the training set is made such that it should resemble the original data quite well, there may be situations in which the range for a given parameter in the dataset is degenerate for all segments, even though it is not in the training set. In such situation, the use of the parameter and its range will not contribute to the evaluation. Consequently, it may be omitted (for performance reasons) or it may be replaced by another a parameter that has a valid range, even if it has worse correlation. Using the original dataset to determine if a parameter does not have a degenerate range for all segments increases the chance of using a good parameter and range. The second benefit is related to performance. The range has to be calculated for the chosen parameters in the given dataset in order to determine the new values. There is no real need to know the range for the training set, so those calculations do not need to be performed.

[1] This might not be the case if the data modelled is very similar in neighbouring cells, as is for instance the case when resampling a grid over a grid with smaller cells. In such situations, the grid gives the impression of a higher accuracy than the data.

6 Experiments

6.1 Prerequisites

In [3], different datasets were considered and tested using three manually and optimally chosen parameters. For the experiments here, a two artificial datasets are considered. These are made from sampling the geometries as shown on Figure 1a and 1b. For these artificial datasets, an ideal solution is known and as such, it is possible to determine what the best parameters are without having to resort to the generation of a test dataset. The geometries are quite different: the first one is comprised of line sources whereas the second one only has area sources. In addition, the reference geometry as described in 5.1 and shown on Figure 1c will be generated. This is a geometry designed to exhibit different properties; it only has line sources, but with a specific pattern and different values for different lines.

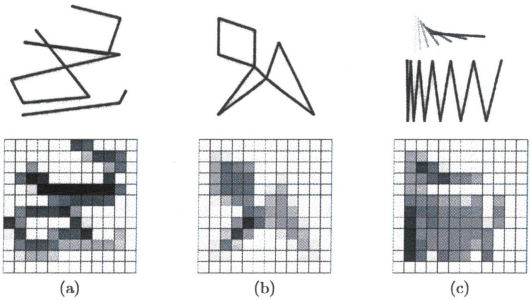

Fig. 1. The three geometries used to test the algorithms and their approximation as input grids. Geometry (a) contains two line sources with a constant value, geometry (b) contains 3 area sources with a constant value and geometry (c) contains different line patters with varying associated values. Greyscales are used to illustrated the values: higher values are shaded darker.

The geometries described above will be considered over two different sets of grids, to generate two different test cases and one reference test case. For both test cases, it is necessary to generate an input grid, an auxiliary grid and an output grid. The input grids for the test cases are also shown on Figure 1. For both cases, the same 12x12 grid will be used for the input.

6.2 Case 1

The target grid for case 1 is shown on Figure 2(a). It is a 25x25 grid that covers exactly the same area as the input grid. The segment grid, obtained from the intersection of input grid and target grid, is also shown on the figure. Figure 2(b-d) shows the approximation of the geometries using the first set of grids for auxiliary grid and segment grid. All grids cover exactly the same area; the auxiliary grid a 15x15 grid and the segment grid used in the calculations. On the figure, different grey scales are used to indicate different ranges of values. The segment grids hold the optimal solution, and show how the data of the input cells should be redistributed.

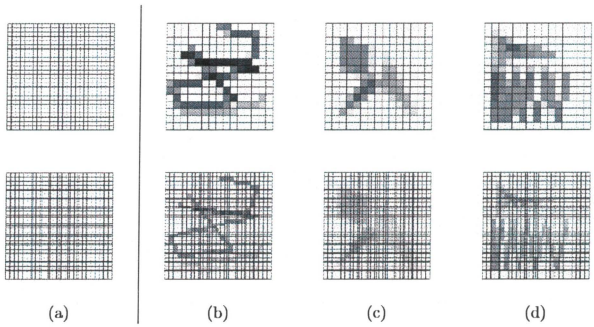

Fig. 2. Test case 1. Target grid (top) and resulting segment grid (bottom) (a). The auxiliary grid (top) and segment grid (bottom), for the first geometry (b), second geometry (c), and reference geometry (d).

All candidate parameters implemented in the prototype were calculated for each output cell. Many of these candidates were dismissed as either having no correlation to the output (Pearson correlation did not yield a number), or as having a degenerate range for all segments. Seven parameters remained. On table 1, the correlations of the seven remaining parameters are listed for the different datasets, in decreasing order of correlation. While the values are different, parameters with the best correlations are similar. The first four parameters occur in the same order. The last three parameters occur in different order, but their correlation is lower than 0.31, which is too low to reliably consider that there is a good correlation.

6.3 Case 2

The target grid for the second case is shown on Figure 3(a). The same 25x25 grid as in the first test case was used, but rotated over 20° counter clockwise. Figure 3(b-d) shows the approximation of the geometries using the second set of grids for auxiliary grid and segment grid. The auxiliary grid is the same 15x15 grid as before, but rotated at a 10° angle compared to the input grid. The calculations

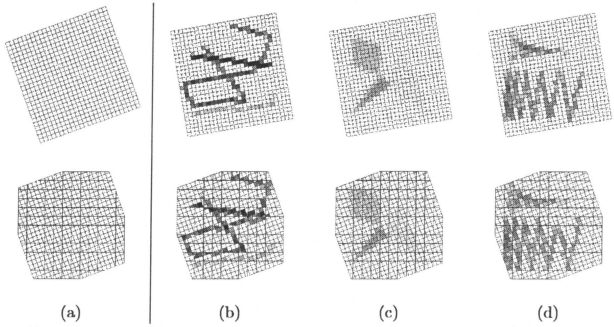

Fig. 3. Test case 2. Target grid (top) and resulting segment grid (bottom) (a). The auxiliary grid (top) and segment grid (bottom), for the first geometry (b), second geometry (c), and reference geometry (d).

were the same as in the previous test case, and again seven parameters remained. The order of the parameters is different compared to the previous test case, even though the approximated geometries are the same: parameters p2 and p3 swapped places; parameter p4 went from being the fourth best to being the worst parameter. This illustrates that the parameters are linked with the way the grids overlap, rather than with the approximated geometries. On table 1, the correlations are listed for the three datasets, in decreasing order of correlation. While the values for the different geometries are different, parameters with the best correlations are similar and occur in the same order.

7 Conclusion and future work

Both testcases show that the choice of which parameters are most suitable is dependent on the layout of the grids. Consequently, it is possible to use a reference data set, which is completely known beforehand, and use the provided rasters to come up with an adequate reference set from which the most suitable parameters for the problem can be determined. This allows for an automatically adjusted

case 1						case2					
dataset 1		dataset 2		dataset 3		dataset 1		dataset 2		dataset 3	
p1	0.65	p1	0.94	p1	0.76	p1	0.76	p1	0.99	p1	0.81
p2	0.60	p2	0.90	p2	0.76	p3	0.76	p3	0.98	p3	0.81
p3	0.56	p3	0.77	p3	0.62	p2	0.64	p2	0.89	p2	0.74
p4	0.30	p4	0.49	p4	0.35	p6	0.32	p6	0.36	p6	0.33
p5	0.18	p5	0.31	p7	0.22	p7	0.20	p7	0.32	p7	0.22
p6	0.13	p7	0.25	p5	0.21	p5	0.18	p5	0.27	p5	0.20
p7	0.12	p6	0.21	p6	0.11	p4	0.03	p4	0.02	p4	0.05

Table 1. The correlations for the parameters in the different datasets in the two considered testcases, in decreasing order of correlation

rulebase system to be generated. The key issue will be determining the reference set. The current choice performs quite well for the current crop of examples, as the order of well correlated parameters is maintained.

In this article, we presented a methodology to find the best suited parameters to use a rulebase system to remap gridded data. The remapping of the data is done by means of auxiliary data that have a known correlation and a rulebase system. The optimal parameters for the rulebase system are found out not dependent on the data, but rather on the grid layouts of the different grids that are involved. This property allows for the generation of a reference data set with an optimal output, from which the optimal parameters can be determined. Using the Pearson correlation, the parameters can be related to the optimal output, and their quality can thus be assessed. The discovered parameters can then be used in the rulebase system to calculate the desired grid transformation. This addition had already been integrated in our current implementation. The current reference geometry is sufficient for the considered problems, but it needs to be studied further if it is universal enough.

References

1. 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.
2. 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.
3. Jörg Verstraete. A fuzzy rulebase aproach to remap gridded spatial data: initial observations. In *IPMU 2014*, page accpeted, 2014.
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. Carol A. Gotway and Linda J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.
6. M Duckham and M. Worboys. An algebraic approach to automated information fusion. *International Journal of Geographic Information Systems*, 19:537–558, 2005.