Environmental Modelling & Software 63 (2015) 87–96

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Effect of spatial data resolution on uncertainty

Mark Pogson [a, b, *], Pete Smith [a]

[a] Institute of Biological & Environmental Sciences, University of Aberdeen, 23 St Machar Drive, Aberdeen AB24 3UU, UK
[b] Engineering, Sports and Sciences Academic Group, University of Bolton, Deane Road, Bolton BL3 5AB, UK

ARTICLE INFO

Article history:
Received 4 March 2014
Received in revised form
12 September 2014
Accepted 25 September 2014
Available online

Keywords:
Uncertainty
Raster grid size
Aggregation
Resolution
Interpolation
Spatial autocorrelation

ABSTRACT

The effect that the resolution of spatial data has on uncertainty is important to many areas of research. In order to understand this better, the effect of changing resolution is considered for a range of data. An estimate is presented for how the average uncertainty of each grid value varies with grid size, which is shown to be in good agreement with observed uncertainties. The effect of bilinear interpolation is also investigated and is observed to provide no reduction in uncertainty relative to uninterpolated data. Finally, the effects of combining aggregated spatial data are found to obey standard properties of error propagation, which means that the presented estimate of uncertainty can be used to estimate resolution-related uncertainty in spatial model results, relative to the input data. The study quantitatively demonstrates the important role of the spatial autocorrelation of data in uncertainties associated with the resolution of spatial data.

1. Introduction

Gridded spatial data are used as parameter inputs and outputs in all kinds of spatial models, including ecological, meteorological and hydrological (Fischer and Wang, 2011; Pogson et al., 2012; Yang et al., 2008). The spatial detail, or resolution, of data affects how well they represent reality, as well as how accurately they can be combined with other spatial data within models to make predictions. The use of gridded data, or rasters, is central to a variety of other disciplines such as biological and medical imaging (Sensen and Hallgrímsson, 2009), hence the effect of data resolution on uncertainty is important for a wide range of research. It is of particular importance to environmental modelling due to the crucial role of spatial data (Beale et al., 2010), the inherent limitations of environmental spatial data (Schimdt et al., 2006), the computational requirements of many models (Wood, 2006), and the large spatial scales often considered (Yue et al., 2011). The present study therefore considers the spatial resolution of data from the perspective of environmental modelling, but the generic formulation is applicable to any field where rescaling and combining of rasters of different resolutions is performed.

Model uncertainty, also referred to as error, can be affected by data resolution not just in terms of the uncertainty of the original data, but also by combining multiple rasters, interpolating data to higher resolutions (Stampfl et al., 2007), and lowering resolution for reasons of computation or compatibility with other rasters. Environmental models commonly take a number of spatial data inputs; for example, a soil organic matter model might require a range of meteorological, land use and soil data (Smith et al., 1997). The model then performs calculations using these data, which amount to different combinations of the data; furthermore, if the data are of different resolutions, there may also be some form of aggregation or interpolation. The way in which uncertainty in the input data is propagated by these operations is crucial to understand the resultant uncertainty in model outputs. The effects of aggregating spatial data are particularly important for spatial optimisation models, which commonly require very low resolution data due to their high computational demands (Wang et al., 2012).

Previous work has investigated the resolution which is appropriate to represent different types of data (Hengl, 2006), as well as the effect that data resolution has on results from specific models (Booji, 2005; Chaubey et al., 2005; Pisoni et al., 2010), the accuracy of particular datasets and derived values (Vaze et al., 2010), error propagation in the production of rasters from observed data (Heuvelink, 1993; Lark, 2000; Knotters et al., 2010), the behaviour of metrics calculated from rasters (Stein et al., 2001) and the level of detail that can reasonably be modelled (Pogson, 2011). The effect of changing the resolution of gridded data has also been considered in

* Corresponding author. Engineering, Sports and Sciences Academic Group, University of Bolton, Deane Road, Bolton BL3 5AB, UK. Tel.: +44(0)1204 903725.
E-mail address: m.pogson@bolton.ac.uk (M. Pogson).

http://dx.doi.org/10.1016/j.envsoft.2014.09.021
1364-8152/

a number of ways, such as by fractal dimension (Bian, 1997). However, the way in which uncertainty changes with the resolution of gridded data has not been predicted for the general case.

This paper considers a range of data operations which are used in spatial models, namely aggregation (lowering resolution), interpolation (increasing resolution) and combination (using a number of datasets to perform calculations). The study first examines the average change in uncertainty of individual grid values caused by aggregating spatial data to lower resolutions, with the aim of predicting the mean uncertainty of values relative to the original, for any grid size. That is, if an aggregated value v' is used to represent a parameter value at point x, the aim is to predict the expected uncertainty of v' relative to that of the original value v at the same point. Consideration of interpolating rasters to higher resolutions and combining aggregated or interpolated rasters is performed in order to see how the uncertainty estimate for aggregating individual rasters can be used to estimate resolution-related uncertainty in model outputs relative to input uncertainty.

Spatial data are considered in this paper as any regular grid of values. While it is uncommon for spatial datasets to be obtained directly from observation, the consideration of uncertainty in the present study is simply relative to the uncertainty of the input data; how the data were obtained or generated does not matter for the present purposes. An estimate is presented for how uncertainty varies with resolution for different types of data distribution, thus enabling prediction of uncertainty for any aggregated grid size. The estimate is tested with a number of artificially generated rasters, as well as data from a widely-used environmental dataset. The effects of interpolating and combining rasters are then investigated by using a number of representative examples. The findings of the study enable quantitative prediction of uncertainties introduced by rescaling gridded spatial data.

2. Methods

2.1. Raster and aggregation definition

We first define an n_1-by-n_2 raster A containing $N = n_1 n_2$ cells; a value n is also defined as the larger of n_1 and n_2. For convenience, each cell in A is identified by a single index i, which is a unique value combining the row index u and column index v. The mean of A is thus:

$$\overline{A} = \frac{1}{N}\sum_{i=1}^{N} A_i \tag{1}$$

The resolution of A is reduced to an n'_1-by-n'_2 raster A', with n' the larger of n'_1 and n'_2, where $n' < n$. This is achieved by aggregating square groups of cells in A (square in terms of number, rather than distance), as commonly performed by a geographical information system (GIS); therefore each cell in A' is an aggregate of g-by-g cells in A (as shown in Fig. 1), and the grid size (the side length of each cell) in A' is g times larger than in A, so $n' = n/g$, where $g > 1$. The present study only presents results for aggregation by the mean of cells (i.e. each cell in A' is the mean of g-by-g cells in A), although use of the mode is also considered in the Discussion. No specific assumption is made of the shape of the grid cells themselves, but the shape must be such that it is not changed by the described aggregation; this is the case for any square or rectangular cells, assuming the grid is regular and aligned. Depending on the raster dimensions and rescaling factor g, some aggregated grid cells at the edge of the raster may include a non-square aggregation of cells, but for a sufficiently large raster the effects of this are not deemed important. Integer values of g are most straightforward to implement but non-integer values are also possible by appropriately weighted allocation of grid cells that span aggregation boundaries.

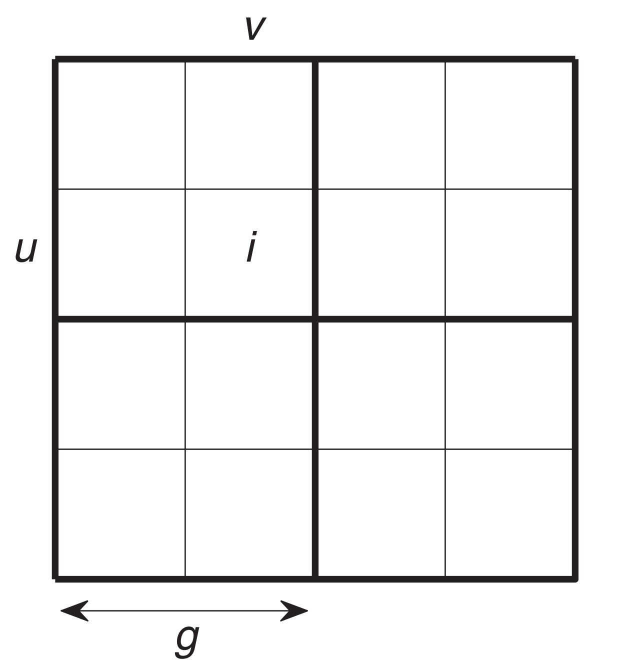

Fig. 1. Grid for original n-by-n raster A (faint lines, $n = 4$), and grid for aggregated n'-by-n' raster A' (heavy lines, $n' = 2$, $g = 2$). For convenience, each cell in A is identified by a single index i, a unique value which combines the row index u and column index v.

In the case of a raster which contains nulls (i.e. values which are either non-numerical or are arbitrary numbers used to signify something other than their numerical value), null values must be excluded from calculations as they cannot be aggregated with other cells (at least, not by aggregation according to the mean). In this case, N is the total number of non-null values in the raster, which means $N < n_1 n_2$ if null values are present.

2.2. Uncertainty definition

Uncertainty E for any aggregated grid size is defined in the present study as the mean absolute error between the original and aggregated raster (Witten and Frank, 2005):

$$E(g) = \frac{1}{N}\sum_{i=1}^{N} |A_i - A'_{i'}| \tag{2}$$

where index i' denotes the cell in raster A' which corresponds to cell i in raster A (hence there are g^2 values of i for each i'). Note that although the described aggregation method is unbiased, $E > 0$ for any aggregated raster (assuming it is not uniform) due to the use of absolute differences, which do not cancel out.

The mean absolute error represents the expected difference between a cell in A and the corresponding cell in A'; it therefore gives the average magnitude of uncertainty caused by using A' instead of A. Mean absolute error is used instead of other metrics, such as mean squared error, as it provides a very tangible measure of error (i.e. the average magnitude of difference between aggregated and original values). However, the choice of metric generally has a relatively straightforward effect on results, as described in the Discussion.

2.3. Spatial autocorrelation

The effect of spatial resolution on uncertainty is clearly dependent on the spatial distribution of values within the raster. For example, if neighbouring cell values are similar to each other, the