Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface

Jin-Feng Wang
LREIS, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, Beijing 100101, PR China; e-mail: wangjf@lreis.ac.cn

Robert Haining
Department of Geography, University of Cambridge, Cambridge CB2 3EN, England; e-mail: rph26@cam.ac.uk

Tie-Jun Liu, Lian-Fa Li, Cheng-Sheng Jiang
LREIS, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, Beijing 100101, PR China; e-mail: liutj@lreis.ac.cn, lilf@lreis.ac.cn, jiangcs@lreis.ac.cn

Received 31 December 2011; in revised form 17 October 2012

Abstract. Spatial sampling is widely used in environmental and social research. In this paper we consider the situation where instead of a single global estimate of the mean of an attribute for an area, estimates are required for each of many geographically defined reporting units (such as counties or grid cells) because their means cannot be assumed to be the same as the global figure. Not only may survey costs greatly increase if sample size has to be a function of the number of reporting units, estimator sampling error tends to be large if the population attribute of each reporting unit can be estimated by using only those samples actually lying inside the unit itself. This study proposes a computationally simple approach to multi-unit reporting by using analysis of variance and incorporating ‘twice-stratified’ statistics. We assume that, although the area is heterogeneous (the mean varies across the area), it can be zoned (or stratified) into homogeneous subareas (the mean is constant within each subarea) and, in addition, that it is possible to acquire prior knowledge about this partition. This zoning of the study area is independent of the reporting units. The zone estimates are transferred to the reporting units. We call the methodology sandwich estimation and we report two contrasting empirical studies to demonstrate the application of the methodology and to compare its performance against some other existing methods for tackling this problem. Our study shows that sandwich estimation performs well against two other frequently used, probabilistic, model-based approaches to multi-unit reporting on stratified heterogeneous surfaces whilst having the advantage of computational simplicity. We suggest those situations where sandwich estimation might be expected to do well.

Keywords: sandwich estimation, heterogeneous surface, zoning; kriging estimates, hierarchical Bayesian estimates

1 Introduction
The provision of statistical estimates of an attribute across many geographically defined reporting units is often required in environmental and social research. Consider the following four examples that demonstrate the different practical contexts in which this can arise. (i) In order to construct pollution maps and to carry out spatially differentiated risk assessments, estimates of heavy metal content in soil are required for each of a large number of 1 km² grid cells or, for example, across each of the 2862 counties of China (Zhou, 2006). (ii) Urban land price mapping is required across a metropolitan area partitioned into a large number of urban tracts. But relevant land price data are available at sample sites where transactions...
have taken place recently (Tsutsumi et al, 2011). (iii) China comprises thirty-four provincial units, 2862 county units, and 41,836 town units, (see http://qhs.mca.gov.cn/article/zlxz/qhtj/200711/20071100003177.shtml). In order to estimate overall tuberculosis (TB) prevalence rates for the country 146 towns are randomly selected from the town units in China (Advisory Panel and Office of 5th National Sampling Survey of Tuberculosis of China, 2010). Guizhou, one of the thirty-four provincial units in China, wants to use the national survey data to estimate its own TB prevalence rate but only five sampled towns are located in Guizhou which is too small a sample size on which to base a reliable estimate. (iv) Flood damage is surveyed by sampling areas close to rivers as well as other areas, but disaster relief resources have to be distributed by administrative units (Xu, 1996). Losses therefore have to be estimated for administrative units using the sample survey data.

The first example involves constructing estimates for a very large number of small geographical areas or reporting units so that overall costs are potentially very high. The second example is a case where only partial data are available on which to base estimates. The third example involves small-sample estimation where, in order to strengthen the estimate for Guizhou (or indeed any other province), there is a need to try to “borrow information” or “borrow strength” from other parts of the sample dataset. The fourth is an example of a problem where statistics collected for one spatial framework have to be transferred to a reporting framework that is quite different. This last problem shares some common ground with the areal interpolation or ‘incompatible areal units’ problem but where the data to be transferred are themselves the outcome of a sampling process (Gotway and Young, 2002). The four examples all have in common the problem of arriving at a good estimate of the mean of an attribute together with an estimate of the sampling error for a set of reporting units which may be very large in number. In all these cases, survey costs will increase greatly if direct sampling has to be employed in which at least two samples have to be taken in each reporting unit, whilst of course the smaller the sample size in a reporting unit the larger the standard error of the estimate of the mean (Cochran, 1977).

In the next section we briefly review existing methods for handling the multi-unit reporting problem and introduce the sandwich estimation approach that will be developed in this paper. Later sections model the propagation of information and uncertainty from what we shall term the sample layer (the geographically distributed sample points) to the zoned layer (the geographical areas on the real surface each assumed to have a constant mean) to the reporting layer (the geographically defined areas for which estimates are required). Our previous study (Wang et al, 2010) investigated how zoning can be employed to improve the estimate of a single global mean, also analyzing the consequences of basing the estimation on an imperfect zoning. This paper develops a novel method for multi-unit reporting (obtaining estimates of the mean which may be different from one reporting unit to another) using a small sample and assuming a zoning that corresponds to the spatial variation in the mean. We then demonstrate the use of sandwich estimation using two contrasting empirical studies and compare findings with other, commonly adopted, approaches to the problem. Finally, we draw conclusions and discuss the implications of this methodology. Readers can implement the two cases presented in the paper and apply the sandwich technique to their own data by visiting http://www.sssampling.org/sandwich.

2 Background and review
Several approaches exist for the multi-unit reporting problem, where the reporting units are many, often small, geographical areas for each of which an estimate of the mean of an attribute (eg, heavy metal content, TB prevalence, flood damage) at some point in time is required (Cochran, 1977, pages 34–39; Särndal et al, 1992, pages 408–412). However, to separate methods it is important at the outset to distinguish between model-based and design-based
approaches to sampling as described in the spatial sampling literature (Brus and de Gruijter, 1997; Christakos, 1992; de Gruijter and ter Braak, 1990; Overton and Stehman, 1995). In design-based approaches the population of values in a region is considered fixed and randomness enters through the process of selecting the locations to sample. The mean value for the region is a fixed but unknown quantity and the sample mean is an estimator of it. Repeated sampling according to a given scheme, such as stratified random sampling, will generate a distribution of estimates of the (regional or population) mean (de Gruijter and ter Braak, 1990). In contrast, the model-based approach assumes the values observed in a region represent one realization of some underlying stochastic model, so observations are treated as random variables with a probability distribution (Cressie, 1993).

The model-based approach to sampling is most appropriate for estimating the parameters of the underlying stochastic model such as its mean or a proportion (Christakos, 1992; Wang et al., 2009), for predicting values at unsampled locations (Matheron, 1963), and certain forms of mapping such as the area-level risk of becoming a victim of burglary (Haining 2003, pages 307–320). The design-based approach is most often used for tackling ‘here and now’ and ‘how much’ questions—estimating global properties such as the mean value of an attribute or the proportion of an area under a particular land use (Cochran, 1977). For a review of these issues see, for example, Haining (2003, pages 96–99).

Universal kriging is a model-based approach to the multi-unit reporting problem. When its assumptions are met it yields the best linear unbiased predictor for attributes distributed continuously in geographical space (Cressie, 1993). Ordinary kriging is a special case of universal kriging under the assumption of second-order stationarity of the attribute where the spatial correlation between two sample points depends only on the distance between them; this assumption implies intrinsic stationarity although the converse is not true (Goovaerts, 1997). Point-to-area kriging can be applied to transfer point estimates to any specified areal framework (Tan et al., 1997), again under the assumption of second-order stationarity. Kriging requires considerable experience to implement, not least in estimating and modeling the variogram which is central to the application of this statistical methodology. More seriously, however, when the spatially varying mean cannot be modeled by a continuous function, even if the area can be partitioned into homogeneous zones and the means subtracted from the sample data in each zone, unless these ‘residuals’ are second-order stationary across the area, kriging has to be implemented on each zone independently which will severely reduce the sample size for the estimation and modeling of the variograms on which the methodology depends (Wang et al., 1997; 2009; 2010).

Another model-based approach to the problem, particularly appropriate to the case of small areas, uses hierarchical Bayesian (HB) modeling (Cressie and Wikle, 2011; Haining, 2003; Rao, 2003; Särndal et al., 1992). This method depends on ‘borrowing information’ (or ‘borrowing strength’) from neighbouring areas. A strategy similar to this has been employed by the US Census Bureau to estimate missing household data and also in the US Medical Expenditure Panel Survey. Estimating area means is based on some specified spatial function that represents the spatial correlation in the observations. There are two potential problems with this approach in the present context: first, there is often a homogeneity assumption—the spatial function is the same across the map, an assumption that may be difficult to sustain across a large study area; second, neighbouring areas are not necessarily the most appropriate areas from which to borrow information.

These model-based approaches assume an underlying probability model generating the observations, so that what is observed is but one realization of the underlying model. It is precisely the willingness to assume a model for the data that makes it possible to obtain estimates for locations or for areas for which no data exist. By contrast, design-based approaches treat observations as fixed quantities, apart from any measurement error (Cressie and Wikle, 2011). For example,
Cochran (1977, pages 142–144) addresses the multi-unit reporting of a zoned surface but in order to calculate means and variances samples are needed in each of the intersected substrata. Särndal et al (1992, chapter 10) also address the problem, but again their solution requires at least one observation in each reporting unit (page 409). Of course, more than one observation is required in each reporting unit so as to provide an estimate of the standard error of the estimate of the mean. In the absence of a model for the observed data, there is no possibility of interpolating to areas where no samples have been taken—hence the need to sample in every reporting unit, preferably taking several samples so as to estimate the standard error but also to improve the estimate of the mean. As the number of reporting units increases, survey costs increase.

In this study, we introduce an approach to stratified estimation of the mean in areas from which no, or too few, sample observations have been taken. The conceptual model of spatial variation which is invoked in order to make estimation possible is referred to as the ‘formal’ (or ‘uniform’) regional model of spatial variation. This model has a long tradition in geography [for an extended discussion see Grigg (1967)], predating the probabilistic models of spatial variation used by statisticians and which underpin the previously described ‘model-based’ approaches to spatial sampling (eg, Cressie, 1993; Ripley, 1981). Unlike statistical models, there is no assumption with the geographers’ regional model that the observations are random variables, so in that sense the approach here is similar to the design-based approach in which observations are treated as fixed. However, the assumption of an underlying model (albeit a model quite different from that usually assumed for model-based approaches to sampling) is critical to the methodology.

Formal regional models partition space into homogeneous or quasi-homogeneous areas (regional “patches”) which represent a classification of space in terms of attribute similarity and spatial contiguity (Haining, 2003, page 183). This model of spatial variation allows the mean for any area (eg, any reporting unit) to be estimated as a function of the means of the homogeneous zones that overlap it.

Versions of this conceptual model of spatial variation arise in a number of areas of spatial analysis, providing a basis for stratified estimation and for some nonparametric solutions to problems including spatial interpolation and areal interpolation (Haining, 2003, pages 131–135; 164–165). The use of this model in the current context allows us to generalize existing spatial estimation theory without the need to assume a probabilistic model of spatial variation of the sort that underpins kriging (which depends on estimating a permissible semivariogram function) or HB estimation (which is based on probability density functions for the data and the prior). We develop what we call sandwich estimation of parameters for multi-unit reporting on heterogeneous surfaces that have been zoned into homogeneous subareas. The procedure consists of two phases: first, the heterogeneous surface is zoned into subareas (partitioned into regions) within which the mean is constant and which provide the framework for spatial sampling; then, two transfer functions are specified. First, the sample data are used to provide estimates of sample means and their variances for each of the zones; next these estimates are transferred onto the multi-unit reporting layer. Of course, the zoning (the model) must provide a good representation of the ‘real’ spatial variability in the mean and there will be bias and a loss of estimator precision in the final estimates when this is not achieved (see Wang et al, 2010). The zoning layer and the reporting layer are two independent partitions of the study area. This model-based strategy allows estimation of the mean and its sampling error even if there is no sample in a reporting unit. We propose this methodology as a simple and direct approach to multi-unit reporting on real surfaces that are heterogeneous in the mean of the attribute but where the mean is constant within each defined zone. The sandwich estimation approach shares common ground with the concept of a layer in a GIS and this model-based assumption, of the presence of regional ‘patches’, is one that has been previously invoked in GIS-based map
operations [see, for example, Flowerdew and Green (1989) and Goodchild et al (1993) in the case of areal interpolation]. For this reason we believe it will be easier for environmental and social scientists to apply, relative to probabilistic model-based approaches, particularly in a GIS computing environment.

3 Sandwich estimation

3.1 Framework for sandwich estimation

Figure 1 provides a pictorial representation of sandwich estimation, which consists of a reporting layer, a zoning layer, and a sampling layer. It is called sandwich estimation because of the three layers of the structure. First, produce a zoning or surface classification that partitions the research area into subareas that are spatially homogeneous (constant mean). Note that any particular class may occur in the form of several geographically separate zones. The purpose of the partitioning is to create distinct zones that subdivide the study area (Wang et al, 2010). Second, distribute sample units over each zone and estimate the sample means and their sampling errors for each of the zones. We recommend stratified random sampling within each zone, wherever possible, as this method of spatial sampling has a long history of application in the geography and spatial statistics literature (see for example Berry and Baker, 1968; Dunn and Harrison, 1993; Hancock, 1995; Overton, 1987; Overton and Stehman, 1993; Ripley, 1981; Wang et al, 2010). Finally, transfer the values from the zoning onto the reporting layer, which consists of many reporting units. Information flows from the sampling layer to the zoning layer and finally to the reporting layer, and comprises estimates of means and their sampling errors.

Figure 1. Conceptual model of sandwich estimation. Note the sampling layer is also a function of the zoning layer which is a product of the object layer.

3.1.1 Reporting layer \( \Psi \)

The reporting layer consists of spatial units. They could be administrative units of a city, counties, postal zones, or census units of a region. They could be a grid system in a soil, ecological or meteorological survey, or physical units such as watersheds or defined by elevation.
3.1.2 Zoning layer \( \{ \mathcal{R} \} \)
Prior knowledge about variability on the real surface can be used either to reduce sample size for a given level of estimator precision or to improve estimator precision (Griffith et al., 1994; Ripley, 1981; Wang et al., 2002; 2010). The importance of taking into account spatial variability and spatial structure in the various stages associated with spatial sampling has been well documented (Anselin, 1988; Griffith, 2005; Haining, 1988; Rodriguez-Iturbe and Media, 1974; Tobler and Kennedy, 1985). Spatial heterogeneity is another feature of spatial variability particularly when sampling over large areas (Foody, 2003; Goodchild and Haining, 2003; Stehman et al., 2003). The purpose of zoning is to partition the area into homogeneous subareas and this may be achieved using quantitative and/or qualitative prior knowledge, about which we will say more later (Li et al, 2008; Wang et al, 1997; 2010).

3.1.3 Sampling layer \( \{ \mathcal{Z} \} \)
In existing design-based sampling (Brus and de Gruijter, 1997; Cochran, 1977; Rao, 2003), whether using simple random, systematic, or stratified random sampling, where sample units are drawn from each reporting unit, the sample size increases in proportion to the number of reporting units: \( \{ \mathcal{Z} \} \propto \{ \mathcal{Y} \} \). In sandwich estimation, sampling is conducted over the zoning layer, so: \( \{ \mathcal{Z} \} \propto \{ \mathcal{R} \} \), which is independent of the reporting layer \( \{ \mathcal{Y} \} \). This strategy cuts the link between reporting layer \( \{ \mathcal{Y} \} \) and sampling layer \( \{ \mathcal{Z} \} \), allowing multi-unit reporting using smaller samples distributed over the zoning layer \( \{ \mathcal{R} \} \). Sampling costs do not increase with the number of reporting units, and users can define reporting units as required for the project, independent of the sampling layer. There is the additional benefit that should there be a need to provide estimates for a different set of areal units than originally planned at the time of the survey, there will be no need to resample; nor, if there is a requirement to provide estimates for several different reporting units at the same time, will this complicate the sampling strategy or add greatly to the cost. Further benefits may arise if sampling involves measuring a wide variety of different properties simultaneously at each site, such as a range of radiological, chemical, and biological properties perhaps in different states (solid, liquid, and gaseous) in order to develop a comprehensive set of maps of different forms of pollution (Staab and Blackhart, 2006). The dual of this multisample problem is where the results of the sample will be used in different ways and for different purposes as arises in the case of certain types of household surveys (see, for example, Fraboni et al, 2005). In all these cases the major consideration is to obtain sufficient samples to provide the desired levels of accuracy and precision required for the estimation of the zone means.

3.2 Implementing sandwich estimation
The objective is to estimate the mean value of an attribute and its standard error for each reporting unit. In sandwich estimation, a sample is drawn from within each of the homogeneous zones denoted \( \{ A, B, C, D \} \) in figure 2. Then the data from the stratified

![Figure 2. Transferring sample estimates from zones to reporting units. The estimate for the reporting unit uses all the sample data collected in A2, B1, C, and D. If B2 is in the same zonal class as B1 then its data may also be used, and the same applies for the data from A1 if A1 is in the same zonal class as A2.](image)
### Table 1. Notation.

<table>
<thead>
<tr>
<th>Property</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>zth zone unit</td>
</tr>
<tr>
<td></td>
<td>sample population</td>
</tr>
<tr>
<td>Number of sample units</td>
<td>( n_z )</td>
</tr>
<tr>
<td>Number of strata</td>
<td>( L = ) number of zones in study area ( \Omega )</td>
</tr>
<tr>
<td>Value of elements</td>
<td>( { y_{zi}, i = 1, \ldots, n_z } )</td>
</tr>
<tr>
<td>Observable total</td>
<td>( y_z \equiv y_1 + \ldots + y_{n_z} )</td>
</tr>
<tr>
<td>Observable mean</td>
<td>( \bar{y}<em>z \equiv \frac{1}{n_z} (y_1 + \ldots + y</em>{n_z}) )</td>
</tr>
<tr>
<td>Variance</td>
<td>( s_z \equiv \sum_{i=1}^{n_z} (y_{zi} - \bar{y}_z)^2 )</td>
</tr>
<tr>
<td>Weighting by size</td>
<td>( W_{z, \text{pop}} = \frac{N_z}{N} )</td>
</tr>
<tr>
<td>Weighting by variance</td>
<td>( W_{z, \text{var}} = \frac{N_S_z}{N_S} )</td>
</tr>
<tr>
<td>Variance of sample mean</td>
<td>( v(\bar{y}_z) = E (\bar{y}_z - E \bar{y}_z)^2 )</td>
</tr>
<tr>
<td>Relative error of mean</td>
<td>( R_z = \frac{</td>
</tr>
<tr>
<td></td>
<td>rth reporting unit</td>
</tr>
<tr>
<td></td>
<td>sample population</td>
</tr>
<tr>
<td>Number of sample units</td>
<td>( n_r )</td>
</tr>
<tr>
<td>Number of strata</td>
<td>( L_r = ) number of zones in ( r ); ( M = ) number of reporting units</td>
</tr>
<tr>
<td>Value of elements</td>
<td>( { y_{ri}, i = 1, \ldots, n_r } )</td>
</tr>
<tr>
<td>Observable total</td>
<td>( Y_r \equiv y_1 + \ldots + y_{n_r} )</td>
</tr>
<tr>
<td>Observable mean</td>
<td>( \bar{Y}<em>r \equiv \frac{1}{N_r} (y_1 + \ldots + y</em>{N_r}) )</td>
</tr>
<tr>
<td>Variance</td>
<td>( s_r \equiv \sum_{i=1}^{n_r} (y_{ri} - \bar{Y}_r)^2 )</td>
</tr>
<tr>
<td>Weighting by size</td>
<td>( W_{r, \text{pop}} = \frac{N_r}{N} )</td>
</tr>
<tr>
<td>Weighting by variance</td>
<td>( W_{r, \text{var}} = \frac{N_S_r}{N_S} )</td>
</tr>
<tr>
<td>Variance of sample mean</td>
<td>( v(\bar{Y}_r) = E (\bar{Y}_r - E \bar{Y}_r)^2 )</td>
</tr>
<tr>
<td>Relative error of mean</td>
<td>( R_r = \frac{</td>
</tr>
</tbody>
</table>

Note: \( \Omega \) refers to the entire area to be surveyed; subscripts \( i, z, r, rz \) refer to sample unit, zoning unit, reporting unit, and zone unit within a reporting unit \( r \), respectively. Lower case letters such as \( y \) and \( n \) refer to sample element, upper case letters such as \( Y \) and \( N \) are population values; \( n_{Sz} \) and \( N_{Sz} \) denote number of sample units and total units in \( \Omega \), respectively; \( n_z \) refers to the number of sample units in zone \( z \) of reporting unit \( r \). An estimated value is denoted by a capped letter.
surface (bottom of figure 2) are transferred to the reporting unit (top of figure 2). Sandwich estimation represents a novel method which in the case of multi-unit reporting avoids the heavy dependence upon intra-area sampling that characterizes conventional design-based methods (Rao, 2003; Särndal et al, 1992) and which can be applied without the parametric assumptions required in the case of probabilistic model-based methods (Christakos, 1992; Matheron, 1963).

The notation used in this study to describe the methodology is listed in table 1.

3.2.1 Distributing sample units over zoning \( \{z\} \)
Sampling is conducted over each zone, resulting in a sample of observations \( \{y_i\}, i = 1,\ldots, n_{\Omega} \). Each sample point has a unique location and the sample refers to a specified time period.

\[
\bar{y}_\Omega = \sum_{i=1}^{L} (W_i \bar{y}_i),
\]

(1)

\[
v(\bar{y}_\Omega) = \sum_{i=1}^{L} W_i^2 v(\bar{y}_i),
\]

(2)

\[
n_\Omega = \sum_{i=1}^{L} n_i,
\]

where, \( \bar{y}_\Omega \) is an estimator of the global population mean \( \bar{Y}_\Omega \), \( \bar{y}_i \) is an estimator of the population mean \( \bar{Y}_z \) of zone \( z \), and \( L \) is the number of zones. \( W_i \) refers to the weight given to the estimate from zone \( z \). There are various ways to distribute the \( n_\Omega \) samples across the zones. One method is to allocate the same number of sample points to each zone (‘same’) irrespective of zone size. Another method, suggested by Cochran [1977, page 93, equation (5.8)], is in proportion (‘proportion’) to the size of zone \( z \) within area \( \Omega \):

\[
\frac{n_{\text{same}}}{n_\Omega} = \frac{N_z}{N_\Omega} = \frac{N_z}{\sum_{i=1}^{L} N_i} = W_{\text{same}}.
\]

(3)

This weighting needs no additional information, resulting in a quick and easy implementation of a distribution [for example \( n_{\text{same}} = n_\Omega (N_z/N_\Omega) \)]. A third method of distributing samples, again see Cochran [1977, page 98, equation (5.26)], is in proportion to the within-zone variance \( \sigma_z \) and the size of the subarea of zone \( z \) as a proportion of the entire area \( \Omega \) (‘variance’):

\[
\frac{n_{\text{var}}}{n_\Omega} = \frac{N_z \sigma_z}{\sum_{i=1}^{L} N_i \sigma_i} = W_{\text{var}}.
\]

(4)

For example, \( n_{\text{var}} = n_\Omega \left( N_z \sigma_z / \sum_{i=1}^{L} N_i \sigma_i \right) \). It is of course necessary to acquire an estimate of \( \sigma_z \)—for example, the sample variance \( s_z \) for each of the zones (see table 1). So, the number of sample units in each zone \( n_z \) could be \( n_{z,\text{same}}, n_{z,\text{var}}, \) or \( n_{z,\text{prop}} \).

3.2.2 Estimating the value for a zone from a sample
After choosing the sample size (\( n_{z,\text{same}}, n_{z,\text{var}}, \) or \( n_{z,\text{prop}} \)) for each zone we estimate the mean and its sample variance for each zone based on the sample data. The sample mean of zone \( z \) is

\[
\bar{y}_z = \frac{1}{n_z} \sum_{i=1}^{n_z} y_{zi},
\]

(5)
which is an unbiased estimator for the design-based population mean $\bar{Y}_r$ with variance $v(\bar{Y}_r)$ where:

$$v(\bar{Y}_r) = E(\bar{Y}_r - E\bar{Y}_r) = \left(\frac{1 - n_z}{N_z}\right)\left(\frac{1}{n_z}\right)s^2_z,$$

$$s^2_z = \left(\frac{1}{n_z}\right)\sum_{i=1}^{n_z} (y_{iz} - \bar{Y}_z)^2.$$  

(6)

(7)

The $\bar{Y}_z$, $s^2_z$, and $v(\bar{Y}_z)$ are calculated for all zones from the sample data, and are then transferred onto the reporting layer.

3.2.3 Estimating values for reporting units from zoning unit estimates

A reporting unit ($r$) may overlap one or several zones ($L_r$). Again, as with stratified sampling [see Cochran, 1977, equation (5.1) on page 91, and equation (5.3) on page 92],

$$\bar{y}_r = \frac{1}{N_z} \sum_{z=1}^{L_z} n_z \bar{Y}_z = \sum_{z=1}^{L_z} W_z \bar{Y}_z,$$

$$v(\bar{y}_r) = \sum_{z=1}^{L_z} W_z^2 v(\bar{Y}_z).$$  

(8)

(9)

Note that the mean and variance of a reporting unit can be estimated even if no sample points occur inside the reporting unit because the quantities for the reporting units are estimated from the overlapping zones, not from the samples inside the reporting unit.

3.3 The philosophy behind sandwich estimation

3.3.1 Borrowing strength from the same attribute class

A stratum of the real surface will often be larger in size than the part that intersects with a reporting unit. It follows that the estimate for the mean of a reporting unit uses not only samples from within it but also from outside it. It is in this way that the sandwich estimator ‘borrows strength’: that is, from all the data obtained from all the zones that overlap a given reporting unit in the sandwich estimation framework—zones which in some cases will extend beyond the boundaries of the reporting unit. Nor does the estimator just borrow strength from neighbouring areas as in the case of model-based Bayesian approaches (Haining, 2003; Rao, 2003; Särndal et al, 1992). This implies that the process of borrowing strength is based on similarity of attribute values across the entire space or some user-defined subset of the space, not merely spatial proximity. The design means that the sandwich estimator has higher precision than conventional statistics (Cochran, 1977), and even works when there are no sample observations at all in a reporting unit, the situation in which existing design-based statistics are unable to provide estimates (Cochran, 1977; Rao, 2003; Särndal et al, 1992).

If the quality of estimates obtained from model-based sampling procedures depends critically on the data meeting distributional assumptions, the sandwich estimation procedure clearly depends critically on constructing a valid zoning of the surface to be sampled. We now turn to this issue.

3.3.2 Zoning the area to be sampled

Constructing a good zoning is critical if sandwich estimation is to provide good-quality estimates at the level of the reporting unit (Li et al, 2008). The strategy for achieving this will depend on the particular circumstances of the application. Wang et al (2010, page 528) refer to three main sources of evidence for zone construction: prior knowledge (which may come from theory or from previous empirical research); presampling; and the use of proxy variables that are believed to be correlated with the variable of interest. The strategy used
for constructing the zones will depend on what information and data are available. Note that if good quantitative data, primary or secondary, historical or contemporary, are available on which to base zoning then region building algorithms can be used to formalize the process (Haining, 2003, pages 199–206). Table 2 summarizes a number of specific cases.

<table>
<thead>
<tr>
<th>No.</th>
<th>Prior knowledge or experience</th>
<th>Historical or secondary data</th>
<th>Zoning strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no</td>
<td>no</td>
<td>Use of small administrative units such as blocks, streets, villages, or towns as zones for bigger administrative units or reporting units.</td>
</tr>
<tr>
<td>2</td>
<td>yes</td>
<td>no</td>
<td>Empirical zoning using prior knowledge and/or experience.</td>
</tr>
<tr>
<td>3</td>
<td>no</td>
<td>yes</td>
<td>Use of unsupervised region-building algorithm augmented where possible by additional data evidence or other knowledge.</td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>yes</td>
<td>Combine prior and/or expert knowledge with formal region-building algorithms.</td>
</tr>
</tbody>
</table>

Table 2. Strategies for constructing zones.

4 Case studies
In this section we demonstrate sandwich estimation on two contrasting datasets using data where reporting unit means are known, because a complete enumeration is available. Having constructed a zoning for a dataset, we sample within each zone (three methods of choosing sample size are considered—same, proportion, and variance) and for each zone we estimate the population mean and sampling error. These estimates are then transferred onto the reporting units using the sandwich estimator. Sampling is repeated and the results averaged across the number of replications.

Case study I is a study of land use where the attribute is spatially continuous. Case study II is a village-based study of per capita income levels. Because the villages form an irregular point distribution, random sampling is used within each zone. For case study I both random and grid-stratified random sampling are used within each zone as it is known that the latter has better properties for estimating the mean of a continuous surface (Haining, 2003, pages 103–106). However, for purposes of comparability, and because numerical differences are evident only at the second or third decimal place, we report here only the results for random sampling for case study I.

For comparison, two other estimation methods are implemented on the same sample data—block kriging and HB estimation. Both are widely used model-based methods. In order to make fair comparison with sandwich estimation both the kriging and HB methods were applied to sample data ‘residuals’: that is, sample data after subtracting the sample mean of the zone within which each data point lies. The sample mean is added back in afterwards. In the case of kriging a single, global, semivariogram model was fitted to these residuals (see earlier comments) and in the case of HB the definition of ‘neighbour’ and the spatial function were the same across the study area. To obtain estimates for reporting units from these two methods, a set of estimates are obtained for each zone and then the value for any reporting unit calculated on the basis of averaging the values lying within it.

The model specifications and parameter values used for kriging were chosen and calibrated using the observed data to obtain the best fit. The empirical semivariogram
Sandwich estimation for multi-unit reporting is estimated using the geostatistical analysis tools in ArcGIS (ESRI Inc.) and the best theoretical semivariogram model chosen by cross-validation (see, for example, Cressie, 1993; Ripley, 1981). Further details on the specific implementation are available from the first author and the codes used to implement them are available online (http://dx.doi.org/10.1068/a44710).

4.1 Case study I: multi-unit reporting of cultivated land using a small sample

The problem is to estimate the proportion of cultivated land in Shandong province, China in 2000 for different reporting units. The data report the proportion of cultivated land extracted from a TM remote sensing image stored in 2 km × 2 km grids of which there are 38,293 in the province. The cultivated land is partitioned into four zones reflecting heterogeneity in the spatial distribution of the proportion of land under cultivation. This partitioning into zones uses data from the 1995 cultivated land-use map in order to reproduce the situation where zoning uses prior knowledge. Zoning is obtained by applying the K-means algorithm (MacQueen, 1967) which implements the principle of maximizing the between-zone variance and minimizing the within-zone variance. The final map divided into four zones was felt to provide a satisfactory zoning by those with first-hand experience of the distribution of Shandong’s cultivated land. Figure 3 shows the 2000 data and the zoning.

The sampling rate is 1%, and the number of sample units in each zone is determined by the three methods: same, proportion, and variance (see subsection 3.2.1). There are three sets of reporting units: 17 administrative units, 4 physical watershed units, and 68 (50 × 50 km²) grid units. The relative error of the sample mean for reporting unit \( r \) is defined as:

\[
R_r = \left| \frac{\bar{y}_r - \bar{Y}}{\bar{Y}} \right|
\]

In the case of the grid reporting framework, only those not clipped by the study area boundary are used in the calculation so this statistic is based on 37 of the 68 units.

Figures 4 and 5 show, for the proportion sample selection method, the average proportion of cultivated land and the relative error for each reporting unit across the three estimation methods. Results are based on 500 repeated samplings in the case of sandwich sampling and kriging, but only 50 in the case of HB because of the time taken to compute each run. In the case of kriging the spherical model, which gave the best fit to the data when compared with other permissible variogram models, was used to describe the second-order spatial variation. In the case of the HB method, after inspecting the residuals a normal model was fitted with conditional spatial autoregressive prior (see, for example, Haining 2003). Figure 6 shows graphs of the average relative error for all three sample size selection methods, for all three reporting frameworks, and all three methods of estimation.

For any given reporting framework and given method of distributing samples, the three estimation methods generate average relative errors that are very similar (note the scales on the three axes in figure 6) with no one estimation method consistently producing the smallest value. For any given reporting framework the differences are also small across the three sample selection methods. The variance method tends to generate the largest average relative errors and the proportion method the smallest, for any given estimation method for the grid and administrative reporting units, but not for the watershed units. As might be expected, the average relative error increases as the number of reporting units increases so that the averages for the grid reporting units (37) are uniformly slightly larger than for the administrative reporting units (17) which in turn are much larger than for the watershed units (4).
Turning to the geography of the area ratio estimates (figure 4) and the average relative errors (figure 5) it appears that in the south-central area of the map, which shows the largest zonal variability (figure 3), all three methods tend to oversmooth area estimates (figure 4) and simulations report some of the largest average relative errors in the case of the grid reporting framework which might be expected to be the most sensitive to underlying zonal variability. It is not apparent that in this particular subarea any one method does significantly

**Figure 3.** (a) Cultivated land in 2000 (shown as the proportion of an area) and (b) the zoning layer obtained from the cultivated land map for 1995 in Shandong Province.
better than the others. It appears that, as the number of reporting units increases, and as the local (small area) heterogeneity of the underlying surface becomes more pronounced, all three methods suffer a similar increase in average relative error.

Figure 4. Map of average proportion of cultivated land under repeated sampling for each reporting unit (proportion method for distributing samples across zones).

Figure 5. Map of average relative error of sample proportions estimated for each type of reporting unit (proportion method for distributing samples across zones).
4.2 Case study II: multi-unit reporting of an economy

This example demonstrates the application of sandwich estimation to social statistics. The aim in this study is to estimate per capita income at the village level for different reporting units in Hing’an, Inner Mongolia Autonomous Region, China which has a population of 1.65 million distributed over an area of 60,000 km² [figure 7(a)]. The set of villages is partitioned into five zones to reflect the heterogeneity of village incomes in Hing’an [figure 7(b)]. This partitioning into zones is based on the expectation that the rural population is richer the closer it lives to a city and the higher the nearest city’s gross domestic product. There are three sets of reporting units: five administrative units, three reporting units based on village height above sea level (elevation), and an artificial grid consisting of twenty-six reporting units. Samples are distributed over the five zones using the three methods described for case study I. The kriging and sandwich estimation methods are repeated 500 times, the HB method 50 times with a 10% sampling rate. The spherical model was again found to provide the best fit for describing second-order spatial variation of the ‘residuals’ for the purpose of kriging estimation and for HB estimation a normal model with a conditional spatial autoregressive prior was fitted to the ‘residuals’.

The data are from a village economic survey in 2007. There are two cities and four counties in Hing’an with 854 villages. Since there are only two villages in Arxan city in Northern Hing’an, Arxan city is excluded from this example. We have 548 valid village survey tables and they are taken as the full dataset to check the error associated with the three estimation methods. Figures 8 and 9 show maps of mean income and the average relative error based on repeated sampling using the proportion sampling method. Figure 10 is a graphical summary of the average relative errors comparable to figure 7. Note that for the grid units we only use the grid squares that are not clipped by the study area boundary which only leaves six grid squares. For any given set of reporting units and method of distributing samples, differences between estimation methods are again very small. The average relative errors do not increase in a straightforward way as the number of reporting units increases from three to five to six. The grid square framework, the most artificial of the three reporting units, produces much bigger average relative errors than the other two (administrative units and elevation). This suggests that average relative error may depend not only on the number of reporting units but also on their configuration.

Turning finally to the geography of the results, all three estimation methods produce similar maps of mean income for any given reporting framework (figure 8) and where local, zonal, heterogeneity is most variable (in the north and central areas). Figure 9 suggests that all three estimators produce similar sized relative errors which become larger as the number of reporting units increases.
Figure 7. (a) Villages with valid survey data and (b) the five-zone map for Hing’an, Inner Mongolia.
Figure 8. Map of the average of the sample mean under repeated sampling for each reporting unit in case study II (proportion method for distributing samples across zones).

Figure 9. Map of the average relative error of estimates of the mean for each reporting unit (proportion method for distributing samples across zones).
5 Conclusion and discussion

Allowing for multi-unit reporting can greatly increase survey costs. Additional problems arise if some reporting units have no sample data, which can arise with traditional design-based approaches whether based on stratified random or systematic sampling. Progress in addressing the multi-unit reporting problem requires modeling assumptions, for without them it is not possible to generate estimates at locations or within areas where no samples have been taken. Small area estimation techniques (eg, HB) and kriging methods are based on probability models where the choice of distribution function and covariance or variogram function are critical to their implementation. Sandwich estimation depends on a different model and a different set of assumptions. The area must be able to be zoned in order to partition it into homogeneous or quasi-homogeneous subareas of constant mean. These zones are then sampled separately and results transferred onto the relevant reporting units. It is this assumption, put into operation where possible using available quantitative and expert knowledge, that underpins sandwich estimation and makes inference possible. Clearly the user needs to be able to construct such a regional framework in a rigorous way and the consequences of different amounts of misalignment (between the ‘real’ strata or zones on the ground and those used to structure the sampling design) have been reported elsewhere (Wang et al, 2010). The empirical studies reported above suggest that if satisfactory regional zones can be constructed then sandwich estimation provides a methodology for multi-unit reporting comparable with probability-based models.

The performance of sandwich estimation is affected by the number of reporting units for which estimates are required, the spatial structure of those reporting units, and the structure of local heterogeneity. In general, sandwich estimation performs less well the more spatially fragmented local heterogeneity is and the more reporting unit estimates that are required—but then so do the other methods against which it has been compared here. Other than that the sampling errors associated with the sandwich estimator will be a function of the usual determinants of uncertainty in sampling (sample size or sample layout including the size of the strata used in the case of stratified random sampling) as well as the extent to which the analyst captures the true underlying heterogeneity and partitions it into homogeneous subareas (Arbia et al, 1998; Ripley, 1981; Rodeghiero and Cescatti, 2008). Sandwich estimation may not bring significant benefits if the surface is only weakly heterogeneous or if the adopted zoning does not correspond well to the true underlying heterogeneity. Therefore, every effort must be made to construct a zoning that reflects well the “true” stratification of the real surface. Zoning may be based on prior knowledge, presampling, an effective proxy variable, or on the distribution of other variables known to affect the value of the attribute of interest (eg, physiography in the case of estimating crop yields or vegetation cover). For an extended

The sandwich estimation framework provides a direct and relatively simple methodology for the multi-unit reporting problem and for transferring data between two polygon systems. It can be easily implemented in a GIS environment because both share the concept of ‘layer’. The sandwich estimation procedure has a number of merits. First, it breaks the linkage between the reporting layer and the zoning layer, distributing sample units over the zoning layer rather than over the reporting layer; consequently, the sample size is independent of the number of reporting units and this allows for considerable flexibility in reporting results across different reporting unit frameworks. We see these features as important benefits when using this methodology. The reporting units can be any spatial polygons of interest, and it is not necessary to guarantee at least two sample units in each of the reporting units, as is usually required with design-based sampling methods (Cochran, 1977). Second, the sandwich estimation framework encourages the use of both quantitative and qualitative prior knowledge to improve the efficiency of sampling. This in turn suggests the need to engage different groups and stakeholders in the process of designing a sample. The ground can be zoned, to form the middle layer for the sandwich estimator, using classification algorithms if quantitative data are available (Li et al, 2008), or by simply hand drawing on maps if only qualitative expert knowledge is available. Third, within each zone the samples may be distributed randomly, systematically, or according to a stratified random layout depending on spatial structure (Dunn and Harrison, 1993) and the ease of implementation (Haining, 2010). Where possible, stratified random sampling within each zone is to be preferred, but the process of zoning may mean the gains over random sampling may not be pronounced. This was found in the two case studies reported here, but again this may reflect the particular circumstances of these examples. Results from these limited comparisons do not provide definitive evidence of the relative merits of the different procedures, but do show that in two contrasting sampling situations sandwich estimation performs well against other, more analytically demanding, approaches to the multi-unit reporting problem.

The evidence here suggests that, whilst sandwich sampling does not outperform kriging and HB methodologies, it does provide estimates of comparable quality. In addition, it offers certain advantages over those two other approaches to multi-unit reporting. First, and perhaps most importantly, it is straightforward to implement and may in some circumstances be easier to explain and justify to policy makers and applied scientists since it is not grounded in two-dimensional stochastic process theory, nor does it raise the problem of how to handle study area boundaries (‘edge effects’) in the estimation procedure. Second, the sandwich estimator has the advantage that it ‘draws strength’ in estimation from other data values based on attribute similarity rather than just spatial proximity.

Whilst the sandwich estimator is presented here as based on a different spatial model from that underlying kriging and HB, it is possible to reconcile it and the two other estimators within a single conceptual model of spatial variation. As Cressie (1993, page 25) remarks, distinguishing between large-scale nonstochastic variation (eg, mean structure) and small-scale stochastic variation (eg, autocorrelation), which are the two dominant elements of spatial variation, involves trade-offs (in respect of model fit and parsimony) as well as personal choice. We suggest that sandwich estimation will be most effective in those situations where the former source of spatial variation is considered to dominate the latter, whilst estimation based on kriging and HB will be most effective where stochastic variation is considered to dominate. The structure of that variation is also likely to be significant (eg, whether mean variation is continuous or patch like; whether second-order variation is the same everywhere on the map) with sandwich estimation working better where mean variation is patch like.
and any second-order variation is nonstationary. These observations may provide a focus for further research into methods for multi-unit reporting.

Acknowledgments. This study was supported by the following grants: NSFC (41271404; 41023010); MOST (2012CB955500); and CAS (XDA05090102). The authors wish to thank three anonymous referees for their insightful comments that have helped us to strengthen the paper.

References
Anselin L, 1988 Spatial Econometrics: Methods and Models (Kluwer Academic, Dordrecht)
Brus D J, de Grujtier J J, 1997, "Random sampling or geostatistical modeling? Choosing between design-based and model-based sampling strategies for soil (with discussion)" Geoderma 80 1–44
Cressie N, 1993 Statistics for Spatial Data (John Wiley, Chichester, Sussex)


Xu F Q, 1996 *Disaster Statistics* (Hunan People Press, Hunan)

Online appendix for:
“Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface”

Jin-Feng Wang, Robert Haining, Tie-Jun Liu, Lian-Fa Li, Cheng-Sheng Jiang

† The original paper to which this appendix refers to can be found at http://dx.doi.org/10.1068/a44710
S1 R block Kriging code for Hingán case example.

library(gstat)
library(foreign)
library(debug)

vgmModel <- vgm(psill=151350, model="Sph", range=139040, nugget=441700)

inputFolder <- "E:/sandwich/samples-xin'anmeng/samples";
outputFolder <- "E:/sandwich/samples-xin'anmeng/samples/result";

fold1 <- c("10")
fold2 <- c("pro","size","var")

WaterData <- read.dbf("E:/sandwich/samples-xin'anmeng/samples/landscape.dbf");
GridData <- read.dbf("E:/sandwich/samples-xin'anmeng/samples/grid.dbf");
AdmiData <- read.dbf("E:/sandwich/samples-xin'anmeng/samples/adm.dbf");

main <- function()
{
  for(f1 in fold1)
  {
    for(f2 in fold2)
    {
      inputFolderUp <- sprintf("%s/%s/%s", inputFolder, f1, f2);
      VecFiles <- list.files(inputFolderUp);
      lenFileCnt <- length(VecFiles);

      pb <- txtProgressBar(min = 1, max = lenFileCnt, initial = 1, style = 3) # progress bar in console

      # store result
      WaterResult <- vector();
      ADMResult <- vector();
      GridResult <- vector();

      # store residual
      WaterAbsResidual <- vector();
      ADMAbsResidual <- vector();
      GridAbsResidual <- vector();

      for(f in 1:lenFileCnt)


```{r}
# progress bar
setTxtProgressBar(pb, f);

currentFile <- sprintf("%s/%s", inputFolderUp, VecFiles[f]);

Sample <- read.csv(currentFile); # sample file
sampleCnt <- nrow(Sample);

subZoneCnt <- 5; # knowledge sub zone count

ZoneSampFactor <- factor(Sample$SculpID);
subZoneMean <- tapply(Sample$ppincome, ZoneSampFactor, mean); # knowledge sub zone mean
subZoneLen <- tapply(Sample$ppincome, ZoneSampFactor, length); # knowledge sub zone size

sampZoneMean <- rep(subZoneMean, subZoneLen); # observed sub zone mean
sampResidual <- Sample$ppincome - sampZoneMean; # observed value minus its corresponding sub zone mean

Sample2 <- cbind(Sample, sampResidual); # add residual to the end field of sample file

# begin Kriging
coordinates(Sample2) = ~x + y

krgPred <- read.csv("E:/sandwich/samples-xin'anmeng/samples/villages.csv"); # the file kriging to predict
coordinates(krgPred) = ~x + y;

kriged = krige(sampResidual ~ 1, Sample2, krgPred, model = vgmModel);

SampleKrg <- cbind(as.data.frame(krgPred), as.data.frame(kriged));

# mean residual for each sub zone
ZoneKrgFactor <- factor(SampleKrg$SculpID);
ZoneResidualMean <- tapply(SampleKrg$var1.pred, ZoneKrgFactor, mean);

# Add mean of sub zone to mean of residual
ZoneTotalMean <- subZoneMean + ZoneResidualMean;

# add mean to each predict point
LenSampleKrg <- nrow(SampleKrg);
```
library(gstat)
library(foreign)
library(debug)

vgmModel <- vgm(psill=151350, model='Sph', range=139040, nugget=441700)

inputFolder <- "E:/sandwich/samples-xin'anmeng/samples";
outputFolder <- "E:/sandwich/samples-xin'anmeng/samples/result";

fold1 <- c("10")
fold2 <- c("pro","size","var")

WaterData <- read.dbf("E:/sandwich/samples-xin'anmeng/samples/landscape.dbf");
GridData <- read.dbf("E:/sandwich/samples-xin'anmeng/samples/grid.dbf");
AdmiData <- read.dbf("E:/sandwich/samples-xin'anmeng/samples/adm.dbf");

main <- function()
{
  for(f1 in fold1)
  {
    for(f2 in fold2)
    {
      inputFolderUp <- sprintf("%s/%s/%s",inputFolder,f1,f2);
      VecFiles <- list.files(inputFolderUp);
      lenFileCnt <- length(VecFiles);

      pb <- txtProgressBar(min = 1, max = lenFileCnt, initial = 1, style = 3) # progress bar in console

      #store result
      WaterResult <- vector();
      ADMResult <- vector();
      GridResult <- vector();

      #store residual
      WaterAbsResidual <- vector();
      ADMAbsResidual <- vector();
      GridAbsResidual <- vector();

      for(f in 1:lenFileCnt)
PredTotalMean <- rep(0,LenSampleKrg)
SampleKrg <- cbind(SampleKrg,PredTotalMean)
for(i in 1:subZoneCnt)
{
    SampleKrg[which(SampleKrg$CuldID == i),"PredTotalMean"] <-
    ZoneTotalMean[i];
}

#Water
WaterFactor <- factor(SampleKrg$landScape);
WaterMean <- tapply(SampleKrg$PredTotalMean,WaterFactor,mean);
WaterResult <- rbind(WaterResult,WaterMean);

WaterResi <- abs(WaterMean - WaterData$MEAN);
WaterAbsResidual <- rbind(WaterAbsResidual,WaterResi);

ADM

ADMFactor <- factor(SampleKrg$ADM);
ADMMean <- tapply(SampleKrg$PredTotalMean,ADMFactor,mean);
ADMRresult <- rbind(ADMRresult,ADMMean);

ADMResi <- abs(ADMMean - AdmiData$MEAN);
ADMAbsResidual <- rbind(ADMAbsResidual,ADMResi);

#Grid

GridFactor <- factor(SampleKrg$Grid);
GridMean <- tapply(SampleKrg$PredTotalMean,GridFactor,mean);
GridResult <- rbind(GridResult,GridMean);

GridResi <- abs(GridMean - GridData$MEAN);
GridAbsResidual <- rbind(GridAbsResidual,GridResi);
}
close(pb);

#output results
outputWater <- sprintf("%s\%s\%s_Water.csv",outputFolder,f1,f2);
outputADM <- sprintf("%s\%s\%s_ADn.csv",outputFolder,f1,f2);
outputGrid <- sprintf("%s\%s\%s_Grid.csv",outputFolder,f1,f2);
write.csv(file = outputWater,WaterResi);
write.csv(file = outputADM,ADMResi);
write.csv(file = outputGrid,GridResi);

#output residual
WaterAbsResidual <- vector(); ADMAbsResidual <- vector(); GridAbsResidual <- vector();

for(f in 1:500)
{
    # progress bar
    setTxtProgressBar(pb, f);

    currentFile <- sprintf("%s/%s", inputFolderUp, VecFiles[f]);

    Sample <- read.csv(currentFile); # sample file
    sampleCnt <- nrow(Sample);

    subZoneCnt <- 5; # knowledge sub zone count

    ZoneSampFactor <- factor(Sample$CulID);
    subZoneMean <- tapply(Sample$ppincome, ZoneSampFactor, mean); # knowledge sub zone mean
    subZoneLen <- tapply(Sample$ppincome, ZoneSampFactor, length); # knowledge sub zone size

    sampZoneMean <- rep(subZoneMean, subZoneLen); # observed sub zone mean
    sampResidual <- Sample$ppincome; # - sampZoneMean; # observed value minus its corresponding sub zone mean

    Sample2 <- cbind(Sample, sampResidual); # add residual to the end field of sample file coordinates(Sample2) = ~x + y

    krgPred <- read.csv("E:/sandwichXAM/samples/villages.csv"); # the file kriginged to predict coordinates(krgPred) = ~x + y;

    O <- rep(NA, 548)

    O[Sample2$id] <- Sample2$sampResidual

    buged <- doWinBUGS(O)

    # begin Kriging

    # kriged = krige(sampResidual ~ 1, Sample2, krgPred, model = vgmModel);

    SampleKrg <- cbind(as.data.frame(krgPred), as.data.frame(buged));
# mean residual for each sub zone
ZoneKrgFactor <- factor(SampleKrg$CuldID);
ZoneResidualMean <- tapply(SampleKrg$buged, ZoneKrgFactor, mean);

# Add mean of sub zone to mean of residual
ZoneTotalMean <- ZoneResidualMean ;# + subZoneMean ;

# add mean to each predict point
LenSampleKrg <- nrow(SampleKrg) ;
PredTotalMean <- rep(0, LenSampleKrg) ;
SampleKrg <- cbind(SampleKrg, PredTotalMean);
for(i in 1:subZoneCnt)
{
    SampleKrg[which(SampleKrg$CuldID == i),"PredTotalMean"] <-
    ZoneTotalMean[i];
}

# Water
WaterFactor <- factor(SampleKrg$landScape);
WaterMean <- tapply(SampleKrg$PredTotalMean, WaterFactor, mean);
WaterResult <- rbind(WaterResult, WaterMean);

WaterResi <- abs(WaterMean - WaterData$MEAN);
WaterAbsResidual <- rbind(WaterAbsResidual, WaterResi);

# ADM
ADMFactor <- factor(SampleKrg$ADM);
ADMMean <- tapply(SampleKrg$PredTotalMean, ADMFactor, mean);
ADMResult <- rbind(ADMResult, ADMMean);

ADMResi <- abs(ADMMean - AdmiData$MEAN);
ADMAbsResidual <- rbind(ADMAbsResidual, ADMResi);

# Grid
GridFactor <- factor(SampleKrg$Grid);
GridMean <- tapply(SampleKrg$PredTotalMean, GridFactor, mean);
GridResult <- rbind(GridResult, GridMean);

GridResi <- abs(GridMean - GridData$MEAN);
GridAbsResidual <- rbind(GridAbsResidual, GridResi);

close(pb);
# output results
outputWater <- sprintf("%s\%s\%s_Water.csv",outputFolder,f1,f2);
outputADM <- sprintf("%s\%s\%sADM.csv",outputFolder,f1,f2);
outputGrid <- sprintf("%s\%s\%s_Grid.csv",outputFolder,f1,f2);

write.csv(file = outputWater,WaterResult);
write.csv(file = outputADM,ADMResult);
write.csv(file = outputGrid,GridResult);

# output residual
outputWaterResidual <- sprintf("%s\%s\%s_Water_Residual.csv",outputFolder,f1,f2);
outputADMResidual <- sprintf("%s\%s\%sADM_Residual.csv",outputFolder,f1,f2);
outputGridResidual <- sprintf("%s\%s\%s_Grid_Residual.csv",outputFolder,f1,f2);

write.csv(file = outputWaterResidual,colMeans(WaterAbsResidual));
write.csv(file = outputADMResidual,colMeans(ADMAbsResidual));
write.csv(file = outputGridResidual,colMeans(GridAbsResidual));

}

}

S3 R code used to call WinBUGS

library(R2WinBUGS)
doWinBUGS <- function()
{

model.file <- file.path("E:/temp/bugs/han_mod.txt")
N <- 548
num <- read.csv("E:/sandwichXAM/hb/neiNum.csv")$neiNum
adj <- read.csv("E:/sandwichXAM/hb/neiAdj.csv")$neiAdj
sumNumNeigh <- 3062
tau <- var(O,na.rm=TRUE)
data <- list("N","num","O", "tau","adj", "sumNumNeigh")

init <- function()
{
list(prec.v = 0.5, beta0 = 0, prec.e = 0.5, v=morm(N,0,0), e=morm(N,0,0))
}

parameters <- c("P")

shandong.sim <- bugs(data, init, parameters, model.file,
n.chains = 1, n.iter = 4000, DIC=FALSE,
bugs.directory = "c:/Program Files/WinBUGS14/",

}
FALSE)
    attach.bugs(shandong.sim)
    O <- colMeans(P)
    detach.bugs()
    return(O)

S4 WinBUGS code used for the Hierarchical Bayesian modelling

model {
    for (i in 1 : N) {
        O[i] ~ dnorm(P[i],tau)
        P[i] <- beta0 + v[i] + e[i]
        e[i]~dnorm(0,prec.e)
    }
    # CAR prior distribution for random effects:
    v[1:N] ~ car.normal(adj[], weights[], num[], prec.v)
    for(k in 1:sumNumNeigh) {
        weights[k] <- 1
    }
    # Other priors:
    beta0 ~ dflat()
    #beta1 ~ dflat()
    prec.v ~ dgamma(0.5, 0.0005)  # prior on precision
    prec.e ~ dgamma(0.5, 0.0005)
    #tau ~ dgamma(0.001, 0.001)
}