WORKING PAPERS

N° TSE-397

March 2013

Spatial reallocation of areal data - a review

Réaffectation spatiale de données surfaciques - revue bibliographique

C.Thomas-Agnan and A. Vanhemsz

Toulouse School of Economics

Spatial reallocation of areal data - a review
Réaffectation spatiale de données surfaciques - revue bibliographique

V. H. Do*
C.Thomas-Agnan†
Toulouse School of Economics (GREMAQ), 21 allée de Brienne 31042 Toulouse, FRANCE
A. Vanhems‡
Université de Toulouse, Toulouse Business School and Toulouse School of Economics

March 3, 2013

Abstract

The analysis of socio-economic data often implies the combination of data bases originating from different administrative sources so that data have been collected on several separate partitions of the zone of interest into administrative units. It is therefore necessary to reallocate the data from the source spatial units to the target spatial units. We propose a review of the literature on statistical methods of spatial reallocation rules (spatial interpolation). Indeed one can distinguish several types of reallocation depending on whether the initial data and the final output are areal data or point data. We concentrate here on the areal-to-areal change of support case when initial and final data have an areal support with a particular attention to disaggregation for continuous data. There are three main types of such techniques: proportional weighting schemes also called dasymetric methods, smoothing techniques and regression based interpolation.

Résumé

L'analyse des données socio-économiques engendre souvent l'usage combiné de plusieurs bases de données venant de sources administratives différentes et pour cette raison correspondant à des partitions différentes de la zone d'intérêt en unités administratives. Il est donc nécessaire de réaffecter les données des zones sources vers des zones cibles qui

*e-mail:huyendvmath@gmail.com
†e-mail: Christine.Thomas@tse-fr.eu
‡e-mail: a.vanhems@esc-toulouse.fr

constituent le support spatial de la base de fusion. Nous proposons dans ce travail une revue de la littérature sur les méthodes statistiques de réaffectation (interpolation spatiale). On en distingue plusieurs types selon que le support des données initiales et le support cible sont de type surfacique ou ponctuel. Nous nous concentrons ici sur le changement de support de surface à surface avec une attention particulière au cas de la désagrégation pour variable continue. Il y a trois grands types de telles techniques : les méthodes dasymétriques, les méthodes de lissage et les méthodes à base de régressions.

Keywords : areal interpolation, spatial disaggregation, pycnophylactic property, change of support, polygon overlay problem.

Mots clefs : interpolation spatiale surfacique, désaggrégation spatiale, propriété pycnophylactique, changement de support, problème de superposition de polygones.

JEL Classification : : C21, C31, C53

1 Motivation

Many administrative agencies nowadays are facing the problem of merging information from different administrative origins collected on several incompatible partitions of the zone of interest into spatial units. Note that an easy way to combine data on incompatible supports is to align them on a common grid. For this reason, the EU directive 'INSPIRE' (2007), INfrastructure for SPatial InfoRmation, states principles to "give infrastructure for spatial information to support community environment policies". One of its objectives is to ensure that "it is possible to combine spatial data and services from different sources across the community in a consistent way and share them between several users and applications" and one requirement is that reference data should "enable merging of data from various sources".

The reasons for the existence of incompatible partitions is a historical lack of coordination between collecting agencies, each using its favorite spatial division. Another origin can be the changes of administrative boundaries through time so that the combination of data from different historical periods results in incompatible spatial supports. The support of spatial data refers to the spatial domain informed by each characteristic. It is often that one needs to combine national census statistics with other sources of data, for example in geomarketing or natural sciences. Other examples of such situations arise when some planification task is undertaken such as where should a new school or store be located and the planners need to transfer census data to their particular catchment areas. Even when it is possible to reaggregate the data from the individual level, this solution is time consuming and expensive and may raise confidentiality problems. An easy way to combine data on several different supports is to align them on a common grid and to reallocate all sources to this single target partition. This option (called "carroyage" in French) is currently being exploited in France at INSEE.

This problem is also referred to as the areal interpolation problem. More generally, the change of support problem may involve point-to-point, area-to-point or point-to-area in-

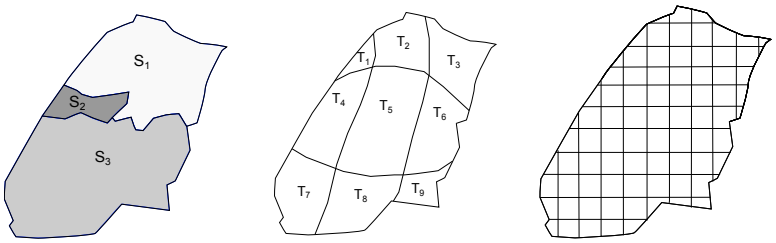

Figure 1: Source zones, target zones and grid target zones.

terpolation. For example, the point interpolation problem is the case of a target variable available for a set of point locations and needed at another location where the data is not available. Gotway and Young (2002) describe these different types and give an overview of the methods. We will focus here on the area-to-area case with a particular emphasis on disaggregation. After introducing the vocabulary and definitions, we will see that there are three main types of such techniques: proportional weighting schemes also called dasymetric methods, smoothing techniques and regression based interpolation. They can be classified by the type of variable they apply to (continuous or discrete, extensive or intensive), the volume preserving property satisfaction (pycnophylactic property), the presence of auxiliary information, the use of simplifying assumptions.

2 Data, vocabulary and aims

The variable of interest that needs to be interpolated is called the **target variable** and it needs to have a meaning on any subregion of the given space. Y_D will denote the value of the target variable on the subregion D of the region interest Ω. In the general area-to area reallocation problem, the original data for the target variable is available for a set of **source zones** that will be denoted by $S_s; s = 1, \cdots, S$ and has to be transferred to an independent set of **target zones** that will be denoted by $T_t; t = 1, \cdots, T$. The variable Y_{S_s} will be denoted by Y_s for simplicity and similarly for Y_{T_t} by Y_t. The source zones and target zones are not necessarily nested and their boundaries do not usually coincide. Figure 1 illustrates these two partitions of the region of interest. With a set of source zones and target zones, one can create a set of doubly indexed intersection zones $A_{s,t} = S_s \cap T_t$, s standing for the index of the source zone and t for that of the target zone. For simplicity, $Y_{A_{s,t}}$ will be denoted by $Y_{s,t}$. Figure 2 illustrates the partition with intersection zones with a zoom on a particular target on the left. Most of the methods will then first proceed to the interpolation of the data from the source to the intersection and in a second step combine the interpolated intersection values to get the target interpolated values. This combination step will require an aggregation rule: one needs to explain how the value of the target variable Y on a zone Ω, Y_Ω, relates to the value of Y on a set of subzones $\Omega_k, k = 1, \cdots p$ forming a partition of Ω. The literature distinguishes between two types

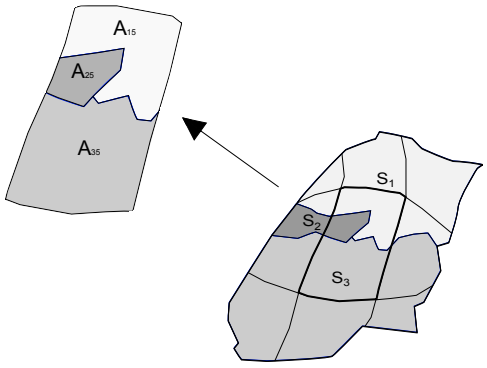

Figure 2: Intersection zones.

of aggregation rules. Let us start with two examples: population and population density. The overall population P_Ω of a region Ω is obtained by simple summation of the population of each subregion P_{Ω_k}. Same is true for any counting variable and such variables are named extensive. Otherwise stated, an extensive variable is a variable which is expected to take half the region's value in each half of the region. Now the population density Y_Ω of the region Ω can be obtained from the densities of the subregions Y_{Ω_k} by a weighted average with weights given by $w_{\Omega_k} = \frac{|\Omega_k|}{|\Omega|}$, where $\mid A \mid$ denotes the area of a subregion A since

$$Y_\Omega = \frac{\sum_k P_{\Omega_k}}{\mid \Omega \mid} = \sum_k \frac{\mid \Omega_k \mid}{\mid \Omega \mid} \frac{P_{\Omega_k}}{\mid \Omega_k \mid} = \sum_{k=1}^{p} w_{\Omega_k} Y_{\Omega_k}.$$

This type of variable is called intensive with weights w_{Ω_k}. More generally linear aggregation takes the general form

$$Y_\Omega = \sum_{k=1}^{p} w_{\Omega_k} Y_{\Omega_k},$$

for a set of weights w_{Ω_k}. If all weights are equal to 1, the variable is called extensive and it is called intensive otherwise. For variables such as population density, we will make use of the following **areal weights matrix**: the (s,t) element of the areal weights matrix W is given by the ratio $w_{s,t} = \frac{|A_{s,t}|}{|S_s|}$ which is the share of the area of source zone s that lies in target zone t. Another example of intensive variable is given by the average price of housing units in a given subregion for a data set of house prices. In this case, the weighting scheme is different and is given by $w_{\Omega_k} = \frac{n_k}{n}$, where n_k is the number of housing units in Ω_k and n is the total number of housing units $n = \sum n_k$. More generally, proportions and rates are intensive variables. Although never really stated, the weights are not allowed to depend upon Y but they may be related to another extensive variable Z by

$$w_{\Omega_k} = \frac{Z_{\Omega_k}}{Z_\Omega}. \tag{1}$$

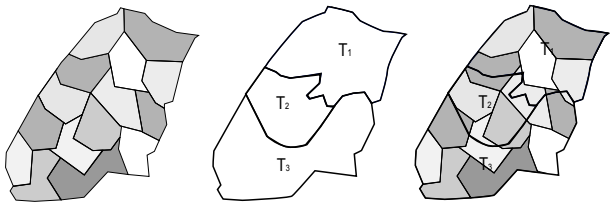

Figure 3: Aggregation case.

In that case note that $w_\Omega = 1$ and that $\sum_k w_{\Omega_k} = 1$. These notions of extensive/intensive variables are also found in physics. Let us note that some variables are neither extensive nor intensive: the target variable Y_A defined by the maximum price on the subregion A is neither extensive nor intensive.
It is important to note that it is always possible to associate an intensive variable to a given extensive variable by the following scheme. If Y is extensive, and if w_A is a weighting scheme of the form (1), the variable

$$\tilde{Y}_A = \frac{Y_A}{Z_A} \tag{2}$$

is intensive since

$$\tilde{Y}_\Omega = \frac{\sum_k Y_{\Omega_k}}{Z_\Omega} = \sum_k \frac{Z_{\Omega_k}}{Z_\Omega} \frac{Y_{\Omega_k}}{Z_{\Omega_k}} = \sum_k w_{\Omega_k} \tilde{Y}_{\Omega_k}.$$

Reversely, if one starts from an intensive variable Y with weighting scheme w_A of the form (1), it can be transformed into an extensive variable by

$$\tilde{Y}_A = Z_A Y_A. \tag{3}$$

Indeed we have

$$\tilde{Y}_\Omega = Z_\Omega Y_\Omega = Z_\Omega \sum_k w_{\Omega_k} Y_{\Omega_k} = \sum_k Z_{\Omega_k} Y_{\Omega_k} = \sum_k \tilde{Y}_{\Omega_k}.$$

Depending on the relative sizes of sources and targets, the areal interpolation problem can be rather of aggregation or disaggregation type. If sources are much smaller in size than targets, one will recover a target value by aggregating sources that will fall inside this target and possibly a few border intersections: this is an aggregation type. In the reverse situation a given target will contain intersections of itself with possible several sources. An intermediate case is when the sizes of sources are comparable to that of targets. Figures 3 and 4 illustrate these cases. We will concentrate here on the disaggregation type.
One property which is often quoted is the so called **pycnophylactic property**. According to Rase (2001), this name comes from the Greek words "pyknos" for mass and "phylax" for guard. This property requires the preservation of the initial data in the following sense: the predicted value on source S_s obtained by aggregating the predicted values on

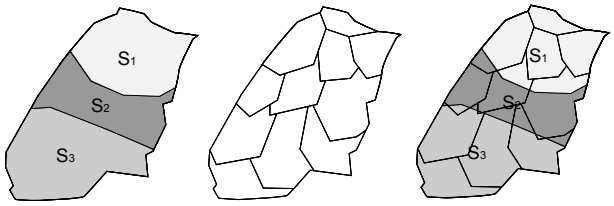

Figure 4: Disaggregation case.

intersections with S_s should coincide with the observed value on S_s. In the case of an extensive variable, this is equivalent to

$$Y_s = \sum_{t:s\cap t\neq\emptyset} \hat{Y}_{s,t}.$$

In the case of an intensive variable with weighting scheme given by w_A, this is equivalent to

$$Y_s = \sum_{t:s\cap t\neq\emptyset} w_{s,t}\hat{Y}_{s,t}.$$

In the literature, one usually encounters this property for the extensive case.
One assumption which is often used to compensate for the absence of information is that of homogeneity. For an extensive target variable, we will say that it is homogeneous in a given zone A if it is evenly distributed within A, meaning that its value on a sub-zone of A is equal to the share of the area of the sub-zone times its value on A. For an intensive variable, we will use the same vocabulary when the variable is constant in each sub-zone of A. The two notions indeed correspond to each other by the relationships (2) and (3).

3 Methods

One early method cannot easily be classified as the others. It is called "point in polygon" and we will describe it first. The others fall into three main classes: the class of dasymetric methods, the class of regression methods and the class of smoothing methods.
Some methods use auxiliary information contained in the observation of an additional related variable to improve the reallocation. When this information is categorical, the level sets of this variable define the so called **control zones**. The spatial support of this auxiliary information can be at the source, target, intersection level or control levels. To expect that the use of X improves the reallocation of Y, we need to believe that Y and X are correlated enough. This raises some questions since Y as well as X are spatial variables hence they can be spatially autocorrelated and it is unclear how to take this into account to correct the classical correlation measures.
Some methods require additional assumptions on the target variable, like for example Y is homogeneous on the sources, or on targets, or the distribution of Y is known to be

Poisson or gaussian.

3.1 Point in polygon

The centroid assignment method also called "point in polygon" allocates the source data Y_s to a target T_t if and only if the source polygon centroid is located within the target polygon. The areal data is thus collapsed to a point datum via a representative point such as the centroid. Voss et al. report that it is the least accurate method. Moreover, it does not satisfy the pycnophylactic property.

3.2 Areal weighting interpolation

It can be applied to an extensive or intensive variable and does not require auxiliary information. For an extensive variable, it is based on the homogeneity assumption that Y_A is proportional to the area $\mid A \mid$. It thus consists in allocating to each subregion a value proportional to the fraction of the area of the source that lies within that subregion. For s such that $s \cap t \neq \emptyset$,

$$\hat{Y}_{s,t} = \frac{\mid A_{s,t} \mid}{\mid S_s \mid} Y_s. \tag{4}$$

After the combination step, this results in the following formula

$$\hat{Y}_t = \sum_{s:s\cap t\neq\emptyset} \frac{\mid A_{s,t} \mid}{\mid S_s \mid} Y_s. \tag{5}$$

For an intensive variable with areal weights, it is based on the assumption that Y is uniform on the sources. It thus consists in allocating to the intersection $A_{s,t}$ the value of Y_s leading to

$$\hat{Y}_{s,t} = Y_s. \tag{6}$$

After the combination step, this results in the following formula

$$\hat{Y}_t = \sum_{s:s\cap t\neq\emptyset} \frac{|A_{s,t}|}{|T_t|} Y_s. \tag{7}$$

It is easy to see that this method does satisfy the pycnophylactic property.

3.3 Dasymetric weighting

Bajat et al. (2011) traces this method back to the 19th century with George Julius Poulett Scrope in 1833 mapping the classes of global population density. The word dasymetry was introduced in the English language by J. K. Wright (1936). The class of dasymetric weighting methods comprises generalizations of areal weighting methods. In order to improve upon areal weighting, the idea is to get rid of the assumption of the count

density being uniform throughout the source zones because this assumption is almost never accurate. For reflecting density variation within source zone, they use other relevant and available information X to distribute Y accordingly. This approach should help allocating Y_s to the small intersection zones within the sources provided the relationship between X and Y be of a proportionality type with a strong enough correlation. Of course it replaces the previous assumption by the assumption that the data is proportional to the auxiliary information on any subregion. This raises the question of how to check the validity of this assumption.
These methods are described in the literature for an extensive variable Y and an extensive auxiliary information X. However it can be adapted to the case of intensive Y as we will see below.
There are some classical examples of auxiliary information for socio-demographic count data or other socio-economic trends coming from road structure or remotely sensed urban land cover data. Yuan et al. (1997) observe a high correlation between population counts and land cover types.
These methods satisfy the pycnophylactic property.

3.3.1 Ordinary dasymetric weighting

It is assumed here that the auxiliary information is known at the intersection level and that it is of a quantitative nature. Voss et al. (1979) propose to use the network of road segments with auxiliary variables like length of roads or number of road nodes to allocate demographic characteristics. Similarly, for target variables such as population or number of housing units, Reibel and Bufalino (2005) use a digital map layer with streets and roads to derive the weighting scheme. The weight of a given subzone is then proportional to the aggregate length of streets and roads in that subzone.
For the case of an extensive target variable with an extensive auxiliary quantitative variable X, the following formulas extend (4) and (5) by substituting X for the area:

$$\hat{Y}_{s,t} = \frac{X_{s,t}}{X_s} Y_s. \tag{8}$$

yielding after the combination step:

$$\hat{Y}_t = \sum_{s:s\cap t\neq\emptyset} \frac{X_{s,t}}{X_s} Y_s. \tag{9}$$

We propose to extend this method to the case of an intensive target variable with weights given by $w_A = \frac{Z_A}{Z_\Omega}$ for a given variable Z and an extensive auxiliary quantitative variable X. The following formulas are obtained using the correspondence intensive-to-extensive given by (2). Indeed, we define the corresponding extensive variable $\tilde{Y}$ by $\tilde{Y}_A = Z_A Y_A$. Using (8) we have

$$\hat{\tilde{Y}}_{s,t} = \frac{X_{s,t}}{X_s} \tilde{Y}_s.$$

We then get

$$\hat{\tilde{Y}}_{s,t} = \frac{X_{s,t}}{X_s} Z_s Y_s,$$

and therefore

$$\hat{Y}_{s,t} = \frac{\hat{\tilde{Y}}_{s,t}}{Z_{s,t}} = \frac{X_{s,t}}{X_s}\frac{Z_s}{Z_{s,t}} Y_s = \frac{X_{s,t}/Z_{s,t}}{X_s/Z_s} Y_s.$$

yielding after the combination step

$$\hat{Y}_t = \sum_{s:s\cap t\neq\emptyset} \frac{Z_{s,t}}{Z_t}\hat{Y}_{s,t} = \sum_{s:s\cap t\neq\emptyset} \frac{Z_{s,t}}{Z_t}\frac{X_{s,t}}{X_s}\frac{Z_s}{Z_{s,t}} Y_s = \sum_{s:s\cap t\neq\emptyset} \frac{X_{s,t}}{X_s}\frac{Z_s}{Z_t} Y_s. \tag{10}$$

3.3.2 Dasymetric weighting with control zones

This is the case when the auxiliary information is categorical, its level sets thus defining the so called control zones. The most classical case, called binary dasymetric mapping, is the case of population estimation when there are two control zones: one which is known to be populated and the other one unpopulated. It is assumed that the count density is uniform throughout control zones. A first step estimates these densities D_c for control zone c by

$$\hat{D}_c = \frac{\sum_s Y_s}{\sum_s \mid S_s \mid},\ s \in c,$$

where $s \in c$ may have several meanings (containment, centroid, percent cover). Assuming now that intersection units are nested within control zones,

$$\hat{Y}_{s,t} = \frac{\mid A_{s,t} \mid \hat{D}_{c(s,t)}}{\sum_{t':s\cap t'\neq\emptyset} \mid A_{t',s} \mid \hat{D}_{c(t',s)}} Y_s,$$

where $c(s,t)$ denotes the control zone which contains the intersection zone $A_{s,t}$. One can see through this formula that this is the same as using ordinary dasymetric with the auxiliary information being a first step crude estimate of variable Y based on the assumption that its corresponding intensive variable (3) is constant throughout control zones. The assumption that intersection units are nested within control zones is not so restrictive since it can be restated as "the control zones are unions of intersections units": control zone information being rather coarse, they can be designed to respect this constraint. Mennis and Hultgren (2006) illustrate this approach with American census data using land cover auxiliary information coming from manual interpretation of aerial photographs.

3.3.3 Two steps dasymetric weighting

This method aims at relieving the constraint of the ordinary dasymetric weighting that the auxiliary information should be known at the intersection level, thus allowing a larger

choice of such information. It is assumed here that the information is known at the level of some control zones which means that the auxiliary information has two components: a quantitative one and a qualitative one. There is a constraint though on the control zones: they should be nested within source zones. The first step is just an ordinary dasymetric step using control zones as targets and the auxiliary information on control zones. Note that the intersection level here is the source-control intersection which is the same as the control level since controls are nested within sources. The second step performs areal weighting with the controls as sources (using the controls estimates of the first step) and the original targets as final targets. Note that the homogeneity assumption used in the second step concerns the control level but since control zones are usually smaller than source zones, the assumption is less constraining. Gregory (2002) presents the implementation of this approach with historical British census data.
If controls are not nested within sources, the method can be adapted by adding an additional step of areal weighting to distribute the control information on the control-source intersections.

3.4 Regression techniques

The dasymetric weighting schemes have several restrictions: the assumption of proportionality of Y and X, the fact that the auxiliary information should be known at intersection level and the limitation to a unique auxiliary variable. The regression techniques will overcome these three constraints. Another characteristic of dasymetric method is that when predicting at the level of the $s-t$ intersection only the areal data Y_s within which the intersection is nested is used for prediction and this will not be the case for regression. In general the regression techniques involve a regression of the source level data of Y on the target or control values of X. The regression without auxiliary information of section 3.4.1 can be regarded as an extension of the areal weighting method since it relies on the "proportionality to area" principle. The regression with control zones of section 3.4.2 is a regression version of the dasymetric weighting with control zones of section 3.3.2. The regression with auxiliary information at target level of section 3.4.3 can be compared to ordinary dasymetric weighting of section 3.3.1.
Generally speaking, these regression methods raise some estimation issues in the sense that very often the target variable is non negative and therefore one would like the corresponding predictions to satisfy this constraint. In order to solve this issue, people resort sometimes to Poisson regression, or ordinary least squares with constraints on the coefficients, or lognormal regression.

3.4.1 Regression without auxiliary information

A first idea discussed in Goodchild et al. (1993) consists in deriving a system of equations linking the known source values Y_s to the unknown target values Y_t using an aggregation formula and an additional assumption of homogeneity of the target variable on the target zones.

In the case of an extensive variable, the homogeneity assumption allows to allocate Y to intersection units proportionally to their area yielding the following system

$$Y_s = \sum_t \hat{Y}_{s,t} = \sum_t \frac{|A_{s,t}|}{|T_t|}\hat{Y}_t$$

For the case of an intensive variable, the homogeneity assumption is that Y is uniform on targets and that its weighting system is given by areal weights. This yields the following relationship between source and target values

$$Y_s = \sum_t \frac{|A_{s,t}|}{|S_s|}\hat{Y}_{s,t} = \sum_t \frac{|A_{s,t}|}{|S_s|}\hat{Y}_t$$

These systems are then solved using an ordinary least squares procedure forced through the origin provided the number of source units is larger than the number of target units. This last condition is not satisfied for disaggregation problems. In that case, one can adapt the technique by combining it with the use of control zones as in section 3.4.2.

3.4.2 Regression with control zones

Using control zones as in section 3.3.2, Goodchild et al. (1993) propose a two steps procedure where the first step is the technique of section 3.4.1 with controls playing the role of targets. The number of such control zones is handled by the user and hence can be forced to be smaller than the number of sources thus relieving the constraint on the number of targets of section 3.4.1. The assumption of homogeneity on targets becomes homogeneity on controls hence it not restrictive because the controls are usually built to reflect homogeneity zones for the target variable. At the end of the first step, one can recover estimates of the target variable at the control level. Using the the uniformity on control assumption, one gets from the control level to the control-target level. The second step in Goodchild et al. (1993) involves a simple aggregation from the control-target intersections level to the target level with homogeneity weights. Yuan et al. (1998) apply rather a dasymetric second step which they call "scaling" using the first step target variable prediction as an auxiliary variable, thus enforcing the pycnophylactic property. Reibel and Agrawal (2006) superimpose a fine grid on the set of source and target zones. They first compute the proportion of each source zone corresponding to each land cover type and then regress the target variable (population) at source level on theses proportions. With the estimated coefficients, they can derive a coarse grid cell based map of the population surface. They rescale these estimates to impose the pycnophylatic property. Then with an aggregation formula they get population estimates for any combination of grid cells, namely for target regions.

3.4.3 Regression with auxiliary information at target level

This family of methods allow to use more than one auxiliary variable and of different natures (quantitative or categorical, or a mixture of both). In Flowerdew et al. (1991),

the emphasis is on extensive target variables with a Poisson or binomial distribution (case 1 hereafter) and in Flowerdew and Green (1992), it is on intensive target variables with a gaussian distribution (case 2 hereafter). In the gaussian case, it is assumed that the target variable Y_A on A is a sample mean of some underlying gaussian variable measured on a number n_A of individuals. Therefore the intensive weights are given by (1) with $Z_A = n_A$ and are approximated by areal weights when the counts n_A are not known. In case 1, we have $Y_{s,t} \sim \mathcal{P}(\mu_{s,t})$, and similarly in case 2 we have $Y_{s,t} \sim \mathcal{N}(\mu_{s,t}, \frac{\sigma^2}{n_{s,t}})$ where the means $\mu_{s,t}$ are in both cases functions of some parameters β and the auxiliary information at target level X_t. In case 2, moreover, it makes sense to assume that $Cov(Y_{s,t}, Y_s) = \sigma^2/n_s$.

With the EM algorithm. Except for a variant in Flowerdew and Green (1992) (see paragraph 3.4.3), the interpolation problem is cast as a missing data problem considering the intersection values of the target variable as unknown and the source values as known therefore allowing to use the EM algorithm to overcome the difficulty.
The algorithm is initialized with areal weighting estimates for $\mu_{s,t}$. The E-step consists in calculating the conditional expectation of $Y_{s,t}$ given the known values Y_s. In case 1, this yields the following formula

$$\mathbb{E}(Y_{s,t} \mid Y_s) = \frac{\mu_{s,t}}{\sum_{t'} \mu_{s,t'}} Y_s$$

which yields the following predictor $\hat{Y}_{s,t} = \frac{\hat{\mu}_{s,t}}{\sum_{t'} \hat{\mu}_{s,t'}} Y_s$ and it is clear that the pycnophylactic property is satisfied.
In case 2, the corresponding formula is

$$\mathbb{E}(Y_{s,t} \mid Y_s) = \mu_{s,t} + \frac{Cov(Y_{s,t}, Y_s)}{Var(Y_s)}(Y_s - \mu_s) = \mu_{s,t} + (Y_s - \mu_s)$$

where μ_s is obtained from the $\mu_{s,t}$ by applying the aggregation formula to the sources subdivided into the intersections and by taking expectation on both sides yielding

$$\mu_s = \mathbb{E}(Y_s) = \mathbb{E}\left(\sum_t \frac{n_{s,t}}{n_s} Y_{s,t}\right) = \sum_t \frac{n_{s,t}}{n_s} \mu_{s,t}. \tag{11}$$

Therefore the E-step yields the following predictor $\hat{Y}_{s,t} = \hat{\mu}_{s,t} + (Y_s - \hat{\mu}_s)$, where the $\hat{\mu}_{s,t}$ come from the previous step and the $\hat{\mu}_s$ from the estimation version of (11).
One can then check that this step enforces the pycnophylactic property since

$$\sum_{t:s\cap t\neq\emptyset} \frac{n_{s,t}}{n_s} \hat{Y}_{s,t} = \sum_{t:s\cap t\neq\emptyset} \frac{n_{s,t}}{n_s} \hat{\mu}_{s,t} + \sum_{t:s\cap t\neq\emptyset} \frac{n_{s,t}}{n_s} (Y_s - \hat{\mu}_s) = \hat{\mu}_s + Y_s - \hat{\mu}_s = Y_s.$$

In the M-step, the intersection values obtained at the previous E-step are considered as i.i.d. observations from the Poisson $\mathcal{P}(\mu_{s,t})$ in case 1 and from the gaussian $\mathcal{N}(\mu_{s,t}, \frac{\sigma^2}{n_{s,t}})$ in case 2. Recall that in both cases, the intersection means are functions of some parameters

β and the auxiliary information at target level X_t plus possibly some information at intersection level such as the area of the intersections. For example in case 1, Flowerdew et al. (1991) consider population as target variable and geology as auxiliary information assuming that the population density will be different in clay areas (λ_1) and in limestone areas (λ_2) so that $\mu_{s,t} = \lambda_t \mid A_{s,t} \mid$, where λ_t is either λ_1 or λ_2 depending on whether target zone t is in the clay or the limestone area. One then performs maximum likelihood with a Poisson regression in case 1 and a weighted least squares in case 2.

Without the EM algorithm. In case 2, Flowerdew and Green (1992) describe a simplified alternative version in the case when one is ready to make the uniform target zone assumption. Namely, since the auxiliary information X is available at target zone level, it does not hurt to assume $\mu_{st} = \mu_t$. Let X_T denotes the $T \times p$ design matrix where p is the number of explanatory factors in X and T the number of targets, μ_S denote the $S \times 1$ vector of source values, μ_T denote the $T \times 1$ vector of target values, W denote the weights matrix whose elements are given by $w_{s,t} = \frac{n_{s,t}}{n_s}$. If we combine the following information:

- the relation between Y and X at target level:

$$\mu_T = X_T \beta,$$

- the aggregation equation $\mu_s = \sum_t \frac{n_{s,t}}{n_s} \mu_{s,t}$
- the uniformity at target level assumption $\mu_{st} = \mu_t$,

we get the following regression equation

$$\mu_S = W X_T \beta \tag{12}$$

between target means at source level and auxiliary information at target level. Using the data at the source level Y_S and equation (12), we can estimate the parameters β by weighted least squares with weights n_s. Then $\hat{\mu}_t = X_t \hat{\beta}$ is a prediction for Y_t.

Alternative with control zone. In case 2, Flowerdew and Green (1992) consider another alternative with a set of control zones assuming that auxiliary information is at control zone level and that it is reasonable to believe that means are uniform on controls $\mu_{s,c} = \mu_c$. The same arguments as above then yield the equations

$$\mu_C = X_C \beta \tag{13}$$
$$\mu_S = W X_C \beta \tag{14}$$

where X_C denotes the $C \times p$ design matrix with C being the number of control zones, μ_C denotes the $C \times 1$ vector of control values, and W being the weight matrix at the source-intersection-control levels. Using the data at the source level and equation (14), we can

estimate the parameters β by weighted least squares with weights n_s. Then $\hat{\mu}_C = X_C\hat{\beta}$ and using the aggregation equation for target and control, one gets that $\hat{Y}_t = \sum_c \frac{n_{c,t}}{n_c}\hat{\mu}_c$ is a prediction for Y_t. Note that one needs two sets of weights $\frac{n_{s,c}}{n_s}$ and $\frac{n_{c,t}}{n_c}$.

3.4.4 Other regression methods

In this section, we describe briefly alternative regression methods. Murakami and Tsutsumi (2011) combine Flowerdew and Green EM algorithm approach with a spatial econometrics regression model to take into account spatial autocorrelation at the intersection unit level. Mugglin and Carlin (1998) propose a hierarchical bayesian version of the Poisson regression method of Flowerdew et al. (1991) with a Markov chain Monte Carlo estimation step and illustrate it on disease counts. The advantage of the hierarchical bayesian approaches is that they provide full posterior distribution estimates enabling accuracy evaluation but their approach requires that the spatial support of the auxiliary information be nested within both targets and source units. Mugglin et al. (2000) extend this approach introducing Markov random field priors on the source and target mean parameters: this allows them to introduce some spatial autocorrelation in the model. They illustrate their approach with population counts reallocation with 39 sources and 160 targets.

3.5 Smoothing techniques

Initially meant for visual display and exploratory analysis, smoothing techniques can solve the point-to-point or the areal-to-point interpolation problems. By laying a fine lattice over the study area and predicting the target variable at each lattice node, they enable mapping the target variable. However they can be used as an intermediate step towards the areal-to-areal interpolation in the sense that once a point prediction is obtained, it is enough to use aggregation rules (integrate the point prediction) to obtain target zones predictions.
In this sense, choropleth mapping is a coarse interpolation technique which amounts, for the intensive variable case, to allocate the areal data value to any point within the support of the corresponding source unit.
Martin (1989) and Bracken and Martin (1991) propose an adaptive kernel density estimation from the target variable values collapsed at the centroids of the source zones. This method is not pycnophylactic.
Tobler (1979) introduces a spline based approach for areal-to-point interpolation. His predictor is a discrete approximation (finite difference algorithm) of the solution to an optimization problem defining a type of interpolating spline with a smoothness criterion based on second partial derivatives. He includes additional constraints such as non-negative point predictions and mass-preservation. His choice of smoothness criterion has been criticized by Dyn et al. (1979). In contrast with Tobler's method which requires a regular grid of prediction points, Rase (2001) adapts Tobler's procedure replacing the

regular grid by a triangulation of the space based on the observed centroids locations, and using some kernel smoothing with inverse distance weighting instead of splines.
Kyriakidis (2004) casts the problem into a geostatistical framework. Indeed the reverse problem of point-to-area interpolation is solved by the block Kriging in geostatistics which is classical due to mining practices: it is of interest for example to predict the total ore content of an area knowing the point data values. Kyriakidis (2004) shows that the area-to-point problem can be solved with similar methods but requires the modeling of all area-to-area and area-to-point covariances. The resulting prediction satisfies the pycnophylactic property. Moreover he proves that choropleth mapping, kernel smoothing and Tobler's pycnophylactic method can be viewed as particular cases of his framework, corresponding to various ways of specifying the covariance model (choropleth mapping corresponding to the absence of correlation at the point support level). A very interesting aspect of the method is that it offers a measure of reliability (standard error of each point prediction). The method can accomodate constraints such as maximum-minimum allowable value or prescribed value of the target variable: for example, zero population value over water bodies or high altitude regions. The method can handle large problems, possibly using moving local neighborhoods. Yoo et al. (2010) adapt it to accomodate more general constraints such as non-negativity. However estimating point covariance from areal data is difficult: it is possible for example with a maximum likelihood procedure based on multivariate gaussian assumption. Liu et al. (2008) propose to combine this approach with regression in an area-to-point residual kriging approach which can be used to disaggregate the regression residuals. Other generalizations can be found in Kelsall and Wakefield (2002) with log-normal kriging.

4 Methods accuracy

The study of the comparative precision of these prediction methods has led to two types of point of views: methodological or empirical.
Unfortunately, there is not much from the methodological point of view since we only found the work of Sadahiro (2000) who considers the point-in-polygon approach and compares it to the areal weighting scheme. He uses a counting process with a fixed number of i.i.d. points with a given density to model the target variable distribution. The target zone is modeled with a fixed shape but a random position. The sources realize a tesselation of the space (considered as unbounded to avoid boundary problems). The last step of the evaluation is of an empirical nature. He finds that the accuracy of point-in-polygon depends upon the target zone size (the bigger the better) and the concentration of the underlying distribution of points. One needs a concentration of points around the representative point in an area of at most 12-15 percent of the total for the point-in-polygon to compare favorably with the areal weighting method, which is quite unrealistic in applications. He also studies the optimal location of representative points which is found to be at the spatial median of the source zone.
The rest of this literature contains many papers of an empirical nature. The comparison of

areal weighting with the alternative dasymetric methods is found in Reibel and Bufalino (2005), Voss (1992), Mennis (2006), Fisher and Langford (1995), Gregory (2002). The dasymetric methods are always found to have better performance than the simple areal weighting with reported improvements up to 71 per cent in relative mean square error (Reibel and Bufalino, 2005).
The comparison of regression methods with several alternatives is found in Flowerdew and Green (1992), Flowerdew et al. (1991), Reibel and Agrawal (2007), Gregory (2002). Flowerdew et al(1991) find that the EM algorithm regression for the Poisson or binomial models performs better than areal weighting by factors of $50-60$ per cent (Poisson case) and $25-55$ per cent (Binomial case) in target deviance. Murakami and Tsutsumi (2011) compare their spatial regression method to more classical regression approaches and find that their spatial lag model performs better. Overall regression methods are found to perform better than dasymetric methods.
For the smoothing methods, Goodchild and Lam (1980) compare areal weighting and Tobler's pycnophylactic interpolation and they do not report any significant advantage for the smoothing method. This may be due to the fact that "count density gradients are not in fact typically smooth up to and beyond tract boundaries" (from Reibel and Agrawal, 2007).

5 Conclusion

We have described the main classes of methods for the area-to-area spatial interpolation problem including proportional weighting schemes also called dasymetric methods, smoothing techniques and regression based interpolation. Among the more sophisticated that we did not give details about, let us mention the multiresolution tree structured autoregressive models (Huang et al. 2002).
In terms of implementation of these methods in usual softwares, there is not much available. Let us mention an implementation of areal weighting in Mapinfo. Bloom et al. (1996) describe their implementation of Flowerdew et al. (1991) with Mapinfo. With R, it is possible to use the "pycno" package by C. Brundson. From our experience with some real data cases, we believe that in large size real applications, the more sophisticated methods are not yet manageable because of size problems and are far too complicated to communicate to the public offices typical users. Simplicity and convenience considerations are certainly the prime arguments for the best choice.
We have not addressed in this review the case of categorical target variable. Indeed in agricultural economics, data are sometimes available at different spatial scales. Chakir (2009) propose a technique for reallocating multinomial type data (namely land use shares) given sampled information at a disaggregated level and observation of aggregated land use shares with a generalized cross-entropy approach.
From the methodological point of view, it is important to note that the only methods including an accuracy measure are are-to-point kriging and the hierarchical bayesian

methods. We think that more attention should be paid to systematic comparisons of the relative accuracies of all these methods in the future.

References

[1] Bajat B., Krunic N. and Kilibarda M., 2011, Dasymetric mapping of spatial distribution of population in Timok Region, proceedings International conference "Professional practice and education in geodesy and related fields, Klavodo-Djerdap, Serbia.

[2] Bloom L.M., Pedler P.J. and Wragg G.E., 1996, "Implementation of enhanced areal interpolation using Mapinfo", Computers and Geosciences, vol. 22, n^0 5, pp. 459-466.

[3] Chakir R., 2009, "Spatial Downscaling of agricultural land-use data: an econometric approach using cross-entropy", Land Economics, vol. 85, n^0 2, pp. 238-251.

[4] Dyn N., Wahba G. and Wong W., 1979, comment on "Smooth pycnophylactic interpolation for geographical regions" by Tobler, Journal of the American Statistical Association, vol. 74, n^0 367, pp. 530-535.

[5] Martin D., Bracken I., 1991, "Techniques for modelling population-related raster databases", Environment and Planning A, vol. 23, n^0 7, pp. 1069-1075

[6] Fisher, P., and Langford, M., 1996, "Modelling sensitivity to accuracy in classified imagery: a study of areal interpolation". The Professional Geographer, 48, pp. 299-309.

[7] Flowerdew R., Green M. and Kehris E., 1991, "Using areal interpolation methods in geographical information systems", Papers in Regional Science, vol. 70, n^0 3, pp. 303-315.

[8] Flowerdew R. and Green M., 1992, "Developments in areal interpolation methods and GIS", The Annals of Regional Science, vol. 26, n^0 1, pp.67-78.

[9] Goodchild M.F., Anselin L., Deichman U., 1993, "A framework for the areal interpolation of socio-economic data", Environment and Planning A, vol. 25, pp. 383-397.

[10] Goodchild M.F., and Lam N., 1980, "Areal Interpolation: A Variant of the Traditional Spatial Problem", Geo- Processing, vol. 1, pp. 297-312.

[11] Gotway C.A. and Young L.J., 2002, "Combining incompatible spatial data", Journal of the American Statistical Association, vol. 97, n^0 458, pp. 632-648.

[12] Langford M., 2007, "Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps", Computers, Environment and Urban Systems, vol. 31, pp. 19-32.

[13] Liu X.H., Kyriakidis P.C. and Goodchild M.F., 2008, "Population-density estimation using regression and area-to-point residual kriging", International Journal of geographical information science, vol. 22, n^0 4, pp. 199-213.

[14] Gregory I.N., 2002, "The accuracy of areal interpolation techniques: standardizing 19th and 20th century census data to allow long-term comparisons", Computers, environments and urban systems, vol. 26, n^0 4, pp. 293-314.

[15] Kelsal J. and Wakefield J., 2002, "Modeling spatial variation in disease risk: a geostatistical approach", Journal of the American Statistical Association, vol 97, pp. 692-701.

[16] Kyriakidis P.C., 2004, "A Geostatistical Framework for Areal-to-Point Spatial Interpolation", Geographical Analysis, vol. 36, n^0 3, pp. 259-289.

[17] Lam N.S., 1982, "An evaluation of areal interpolation methods",Proceedings, Fifth International Symposium on Computer-Assisted Cartography (AutoCarto 5), vol. 2, pp. 471-479.

[18] Martin D., 1989, "Mapping population data from zone centroid locations", Transactions of the Institute of British Geographers, New Series, vol. 14, n^0 1, pp. 90-97.

[19] Mugglin A.S. and Carlin B.P., 1998, "Hierarchical modeling in geographic information systems: ppulation interpolation over incompatible zones", Journal of Agricultural, Biological and Evironmental Statistics, vol. 3, n^0 2, pp. 111-130.

[20] Mugglin A.S., Carlin B.P. and Gelfand A. E., 2000, " Fully model-based approaches for spatially misaligned data", Journal of the American Statistical Association, 2000, vol. 95, n^0 451, pp. 877-887.

[21] Murakami D. and Tsutsumi M., 2011, "A New Areal Interpolation Technique Based on Spatial Econometrics", Procedia-Social and Behavioral Sciences, vol. 21, pp. 230-239.

[22] Rase W., 2001, "Volume-preserving interpolation of a smooth surface from polygon-related data", Journal of Geographycal Systems, vol. 3, pp. 199-213.

[23] Reibel M. and Bufalino M.E., 2005, "Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems", Environment and Planning A, vol. 37, n^0 1, pp. 127-139.

[24] Reibel M. and Agrawal A., 2007, " Areal interpolation of population counts using pre-classified land coer data", Population Research and Policy Review, vol. 26, n^0 5-6, pp. 619-633.

[25] Sadahiro Y., 2000, "Accuracy of count data estimated by the point-in-polygon method", Geographical Analysis, vol. 32, n^0 1, pp. 64-89.

[26] Tobler W.R., 1979, "Smooth pycnophylactic interpolation for geographical regions", Journal of the American Statistical Association, vol. 74, n^0 367, pp. 519-530.

[27] Voss P.R., Long D.D., Hammer R.B., 1999, "When census geography doesn't work: using ancillary information to improve the spatial interpolation of demographic data", CDE working paper, n^0 99-26, Wisconsin, Madison.

[28] Wright J.K., 1936. "A method of mapping densities of population with Cape Cod as an example". Geographical Review, vol. 26, n^0 1, pp. 103-110.

[29] " Yuan, Y., Smith, R. M., and Limp, W. F., 1997, "Remodeling census population with spatial information from Landsat TM imagery. Computers, Environment and Urban Systems", vol. 21, n^0 3, pp. 245-258.

[30] Yoo E., Kyriakidis P.C. and Tobler W., 2010, "Reconstructing population density surfaces from areal data: a comparison of Tobler's pycnophylactic interpolation method and area-o-point kriging", Geographical Analysis, vol. 42, n^0 1, pp. 78-98.

[31] Yoo E., Kyriakidis P.C., 2006, "Area-to-point kriging with inequality-type data", Journal of geographical systems, vol. 8, n^0 4, pp. 357-390.