

“Spatial reallocation of areal data – another look at basic methods

Réaffectation spatiale de données surfaciques - une autre approche des méthodes de base”

V. H. Do, C. Thomas-Agnan, A. Vanhems

Spatial reallocation of areal data - another  
look at basic methods  
Réaffectation spatiale de données surfaciques -  
une autre approche des méthodes de base

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

March 11, 2014

---

\*e-mail: huyendvmath@gmail.com

†e-mail: Christine.Thomas@tse-fr.eu, corresponding author

‡e-mail: a.vanhems@esc-toulouse.fr

## 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 the simplest statistical methods of spatial reallocation rules (spatial interpolation). 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 numerical data. There are three main types of such techniques: proportional weighting schemes also called dasymetric methods, smoothing techniques and regression based interpolation. We propose a unified formalization of the basic techniques with a synoptic table and extensions of some of these methods to new cases.

## 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. Nous proposons dans ce travail une revue de la littérature sur les méthodes statistiques basiques de réaffectation (interpolation spatiale). 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. Nous proposons une formulation unifiée des diverses techniques de base avec un tableau synoptique et des extensions de certaines d'entre elles à de nouvelles situations.

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

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

*JEL Classification* : : C21, C31, C53

## 1 Introduction

The origin of this work is in a collaboration with a French administration, the Midi-Pyrénées DREAL (Direction Régionale Environnement Aménagement Logement) about the merge of several administrative data bases with different spatial support. It was necessary for example to disaggregate the number of housing units, originally available at the commune level, on a fine regular square grid. Similarly, 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. 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 interpolation. 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. A discussion of some methods relative to this framework can also be found in Goodchild et al. (1993) but we go one step further in the degree of formalization and unification.

After introducing the vocabulary and definitions in section 2, we will see that there are three main types of such techniques in section 3. The first type is the family of proportional weighting schemes, also called dasymetric methods, which are illustrated in Yuan et al.(1997), Voss et al. (1979), Reibel and Bufalino (2005), Mennis and Hultgren (2006) and Gregory (2002). The second type is made of regression based interpolation and can be found in Flowerdew et al. (1991), Godchild et al. (1993), Flowerdew and Green (1992) for the simplest ones. The third type comprises smoothing techniques which are described for example in Tobler (1979), Martin (1989), Bracken and Martin (1991), Rase (2001) and Kyriakidis (2004). The set of methods can be classified by the type of variable they apply to (continuous or discrete, extensive or intensive), the volume preserving property satisfaction (pynophylactic property), the presence of auxiliary information, and the use of simplifying assumptions. We could not provide in a single paper the same level of details

for the simple methods and for more complex ones so that we decided to concentrate on the simple ones which are the ones more likely to be adopted by practitioners and to just give some of the main references for the more complex methods. We use a simulated toy example to illustrate some of the methods. In order to ease the practitioner's choice, we present a synoptic table (Table 1) to summarize this classification. We believe that presenting the methods in such a unified way can help the practitioners clarifying the relationships between the very diverse presentations found in the literature. Note that a more detailed and lengthy presentation for practitioners has been written for the DREAL (Vignes et al., 2013). This work of clarification also lead us to find extensions of some of these methods to new cases: for example in section 3.2.1, we extend the ordinary dasy-metric weighting method to the case of an intensive target variable  $Y$  and an extensive auxiliary quantitative variable  $X$  and in section 3.2.2 we show that the assumption of intersection units nested within control zones is unnecessary. Finally, this approach helped us laying the groundwork for a future mathematical evaluation of the respective accuracy of the methods. A brief point on the current state of the literature about accuracy and software issues is done in the conclusion.

## 2 Data, definitions and notations

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 of interest  $\Omega$ . We restrict attention to the case of quantitative target variables (see for example Chakir (2009) for the case of categorical target variables).

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  $S_s; s = 1, \dots, S$  and has to be transferred to an independent set of **target zones** that will be denoted by  $T_t; t = 1, \dots, T$ . 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. Many methods involve the areas of different subregions (sources, targets or other). We will denote by  $|A|$  the area of any subregion  $A$ .

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  $\Omega$ ,  $Y_\Omega$ , relates to the value of  $Y$  on a set of subzones  $\Omega_k, k = 1, \dots, p$  forming a partition of  $\Omega$ . The literature distinguishes between two types of aggregation rules. Let us start with two

examples: population and population density. The overall population  $P_\Omega$  of a region  $\Omega$  is obtained by simple summation of the population of each subregion  $P_{\Omega_k}$ . 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  $Y_\Omega$  of the region  $\Omega$  can be obtained from the densities of the subregions  $Y_{\Omega_k}$  by a weighted average with weights given by  $w_{\Omega_k} = \frac{|\Omega_k|}{|\Omega|}$ , since

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

This type of variable is called **intensive** with weights  $w_{\Omega_k}$ . 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  $w_{\Omega_k}$ . 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  $\Omega_k$  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

$$w_{\Omega_k} = \frac{Z_{\Omega_k}}{Z_\Omega}. \quad (1)$$

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. 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.

Let us show 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

$$\tilde{Y}_A = \frac{Y_A}{Z_A} \quad (2)$$

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

$$\tilde{Y}_A = Z_A Y_A. \quad (3)$$

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 **pyncnophylactic 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 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).

Let us introduce the toy example that will be used to demonstrate some properties. On Figure 5, we can see a square divided into 25 equal cells and three source regions made of unions of cells. The Figure present the values of an auxiliary variable  $X$  in the center panel and the values of two target variables  $Y_1$  on the left and  $Y_2$  on the right. We can see that there is inhomogeneity within sources. The target zones are visible on Figures 6 through 9 which compare some methods through the targets prediction errors.

### 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 dasymmetric

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  $X$  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. We start with the most elementary methods requiring no additional information and complexify progressively.

### 3.1 Elementary methods

#### 3.1.1 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.

#### 3.1.2 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  $|A|$ . 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  $s \cap t \neq \emptyset$ ,

$$\hat{Y}_{s,t} = \frac{|A_{s,t}|}{|S_s|} Y_s. \quad (4)$$

After the combination step, this results in the following formula

$$\hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \frac{|A_{s,t}|}{|S_s|} Y_s. \quad (5)$$

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

$$\hat{Y}_{s,t} = Y_s. \quad (6)$$

After the combination step, this results in the following formula

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

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

From now on, all subsequent methods require additional auxiliary information except in section 3.3.1.

## 3.2 Dasymetric weighting

Bajat et al. (2011) trace 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.

### 3.2.1 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. It might seem difficult to find auxiliary information at intersection level but the following example should convince the user that it is possible. Voss et al. (1979) and Reibel and Bufalino (2005) propose to use the network of road segments with auxiliary variables like length of roads or number of road nodes to allocate demographic characteristics such as population or number of housing units,. 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 formulae extend (4) and (5) by substituting  $X$  for the area:

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

yielding after the combination step:

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

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  $X$ . We define the corresponding extensive variables  $\tilde{Y}$  and intensive variable  $\tilde{X}$  by introducing the transformations from intensive to extensive  $\tilde{Y}_A = Z_A Y_A$  and from extensive to intensive  $\tilde{X}_A = \frac{X_A}{Z_A}$ . The following formula is obtained using the correspondence intensive-to-extensive given by (2) (see the annex for a proof).

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

Similar formulae can be obtained easily in the case  $Y$  extensive with  $X$  intensive and  $Y$  intensive with  $X$  intensive.

Let us illustrate this method with the toy example introduced at the end of section 2. Figure 6 presents a comparison between the results of the areal weighting method and the dasymetric method for target variable  $Y_1$ . Figure 7 does the same for target variable  $Y_2$ . In each target we can see the true value of  $Y_1$  (left) and the value of the prediction (right) and the relative prediction error below  $(\frac{\hat{Y}_1 - Y_1}{Y_1})$ . We can see that the dasymetric method yields better results than areal weighting for variable  $Y_2$  because of the inhomogeneity within sources (indeed the sum of squared errors is 10 percent smaller for dasymetric). However for variable  $Y_1$ , for which the level of inhomogeneity within sources is not as high, this is not the case and areal weighting is doing better than dasymetric with a ratio of sum of squared errors of 48 percent.

### 3.2.2 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

$$\hat{D}_c = \frac{\sum_{s \in c} Y_s}{\sum_{s \in c} |S_s|},$$

where  $s \in c$  may have several meanings (containment, centroid, percent cover). For this method, it is often assumed in the literature that intersection units are nested within control zones in which case the intersection zone prediction is given by

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

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. However let us prove that this assumption is unnecessary. Indeed if one denotes by  $A_{s,t,c}$  the intersection between source zone  $s$ , target zone  $t$  and control zone  $c$ , the following gives a prediction for the target values

$$\hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \frac{\sum_c |A_{s,t,c}| \hat{D}_c}{\sum_{t'} \sum_c |A_{s,t',c}| \hat{D}_c} Y_s.$$

Mennis and Hultgren (2006) illustrate this approach with American census data using land cover auxiliary information coming from manual interpretation of aerial photographs.

### 3.2.3 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. In this case, the intersection level 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. 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.

### 3.3 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 (exceptionally two in the case of two steps dasymetric). The regression techniques will overcome these three constraints. Another characteristic of dasymetric method is that when predicting at the level of the  $A_{s,t}$  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 3.3.1 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 3.3.2 is a regression version of the dasymetric weighting with control zones of section 3.2.2. The regression with auxiliary information at target level of section 3.3.3 can be compared to ordinary dasymetric weighting of section 3.2.1.

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 (as in Flowerdew et al., 1991), or ordinary least squares with constraints on the coefficients (see Goodchild et al., 1993), or lognormal regression (see Goodchild et al., 1993).

#### 3.3.1 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 3.3.2.

### 3.3.2 Regression with control zones

Using control zones as in section 3.2.2, Goodchild et al. (1993) propose a two steps procedure where the first step is the technique of section 3.3.1 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 3.3.1. 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.

### 3.3.3 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  $\beta$  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 3.3.3), 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

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

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

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

where  $\mu_s$  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

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

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  $\beta$  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 ( $\lambda_1$ ) and in limestone areas ( $\lambda_2$ ) so that  $\mu_{s,t} = \lambda_t | A_{s,t} |$ , where  $\lambda_t$  is either  $\lambda_1$  or  $\lambda_2$  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,  $\mu_S$  denote the  $S \times 1$  vector of source values,  $\mu_T$  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  $\beta$  by weighted least squares with weights  $n_s$ . Then  $\hat{\mu}_t = X_t \hat{\beta}$  is a prediction for  $Y_t$ .

Let us consider again the toy example defined earlier to illustrate this technique adapted to the case of Poisson regression. Figure 8 (resp 9) compares the results of this regression technique with the dasymetric method based on the same auxiliary information for  $Y_1$  (resp  $Y_2$ ). For  $Y_1$ , the regression method is better than the dasymetric with a ratio of sum of squared errors of 12 percent. For  $Y_2$  however, the dasymetric is better than the regression with a ratio of sum of squared errors of 82 percent. The reason is that indeed the variable  $Y_2$  has been constructed to be almost proportional to  $X$  (which is in line with the spirit of dasymetric) whereas  $Y_1$  is not. Note that the dasymetric method uses more information than the regression method because it uses the auxiliary value at intersection level whereas the regression method uses it at target level.

**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,  $\mu_C$  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  $\beta$  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 A short overview of more elaborate methods

### 3.4.1 Other regression methods

In this section, we briefly describe alternative regression methods. A detailed development of these more sophisticated techniques would require much more tools and notations. Because one of our objectives is to give priority to the practitioner point of view, we do not develop them in this presentation but just give some of the main references. 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. Huang et al. (2002) introduce multiresolution tree structured autoregressive models.

### 3.4.2 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-area 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. A similar kernel based method is described in Grasland et al. (2000) with a discussion of the relationship between the choice of the bandwidth parameter and the level of aggregation of the initial information.

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 accommodate 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 accommodate 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 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. As we pointed out in the introduction, we have focused on the basic methods which are more likely to be adopted by practitioners, and a summary of the main characteristics of these methods can be found in Table 1.

We have not addressed in this review the case of categorical target variable. 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.

In terms of implementation of these methods in usual softwares, there is not much available. Bloom et al. (1996) describe their implementation of areal weighting from 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.

As mentioned in the introduction, one motivation of this paper was to be a first step for a further study of the comparative precision of these prediction methods. Let us briefly summarize what can be found in the literature so far. Overall one finds 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 tiling partition of the space with geometric shapes (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).

Finally, it is important to point out that the only methods that come along with an accu-

racy measure are area-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.

## 5 Annex: Ordinary dasymetric weighting for intensive target variable and extensive auxiliary variable

For an intensive target variable with weights given by  $w_A = \frac{Z_A}{Z_\Omega}$  and an extensive auxiliary quantitative variable  $X$ , we define the corresponding extensive variable  $\tilde{Y}$  using (3) by  $\tilde{Y}_A = Z_A Y_A$  and intensive variable  $\tilde{X}_A = \frac{X_A}{Z_A}$ . Using (8) we have

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

and therefore

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

which is similar to (8). After the combination step, we get

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

**Acknowledgments.** We would like to thank A. Ruiz-Gazen and C. Vignes for fruitful discussions. This work was supported by the French Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005).

## References

- [1] BAJAT B., KRUNIC N. and KILIBARDA M., 2011, “Dasymetric mapping of spatial distribution of population in Timok Region”, proceedings of 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<sup>o</sup> 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<sup>o</sup> 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<sup>o</sup> 367, pp. 530-535.

- [5] FISHER, P., and LANGFORD, M., 1996, "Modelling sensitivity to accuracy in classified imagery: a study of areal interpolation". *The Professional Geographer*, vol. 48, pp. 299-309.
- [6] FLOWERDEW R., GREEN M. and KEHRIS E., 1991, "Using areal interpolation methods in geographical information systems", *Papers in Regional Science*, vol. 70, n<sup>o</sup> 3, pp. 303-315.
- [7] FLOWERDEW R. and GREEN M., 1992, "Developments in areal interpolation methods and GIS", *The Annals of Regional Science*, vol. 26, n<sup>o</sup> 1, pp.67-78.
- [8] 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.
- [9] GOODCHILD M.F., and LAM N., 1980, "Areal Interpolation: A Variant of the Traditional Spatial Problem", *Geo- Processing*, vol. 1, pp. 297-312.
- [10] GOTWAY C.A. and YOUNG L.J., 2002, "Combining incompatible spatial data", *Journal of the American Statistical Association*, vol. 97, n<sup>o</sup> 458, pp. 632-648.
- [11] GRASLAND C., MATHIAN H. and VINCENT J-M., 2000, "Multiscalar analysis and map generalisation of discrete social phenomena: Statistical problems and political consequences", *Statistical Journal of the United Nations ECE*, IOS Press, vol. 17, pp. 157-188.
- [12] 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<sup>o</sup> 4, pp. 293-314.
- [13] 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.
- [14] KYRIAKIDIS P.C., 2004, "A Geostatistical Framework for Areal-to-Point Spatial Interpolation", *Geographical Analysis*, vol. 36, n<sup>o</sup> 3, pp. 259-289.
- [15] 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.
- [16] 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.

- [17] 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<sup>o</sup> 4, pp. 199-213.
- [18] MARTIN D., 1989, "Mapping population data from zone centroid locations", *Transactions of the Institute of British Geographers, New Series*, vol. 14, n<sup>o</sup> 1, pp. 90-97.
- [19] MARTIN D., BRACKEN I., 1991, "Techniques for modelling population-related raster databases", *Environment and Planning A*, vol. 23, n<sup>o</sup> 7, pp. 1069-1075
- [20] MUGGLIN A.S. and CARLIN B.P., 1998, "Hierarchical modeling in geographic information systems: population interpolation over incompatible zones", *Journal of Agricultural, Biological and Environmental Statistics*, vol. 3, n<sup>o</sup> 2, pp. 111-130.
- [21] 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<sup>o</sup> 451, pp. 877-887.
- [22] 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.
- [23] RASE W., 2001, "Volume-preserving interpolation of a smooth surface from polygon-related data", *Journal of Geographical Systems*, vol. 3, pp. 199-213.
- [24] 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<sup>o</sup> 1, pp. 127-139.
- [25] REIBEL M. and AGRAWAL A., 2007, "Areal interpolation of population counts using pre-classified land cover data", *Population Research and Policy Review*, vol. 26, n<sup>o</sup> 5-6, pp. 619-633.
- [26] SADAHIRO Y., 2000, "Accuracy of count data estimated by the point-in-polygon method", *Geographical Analysis*, vol. 32, n<sup>o</sup> 1, pp. 64-89.
- [27] TOBLER W.R., 1979, "Smooth pycnophylactic interpolation for geographical regions", *Journal of the American Statistical Association*, vol. 74, n<sup>o</sup> 367, pp. 519-530.
- [28] VIGNES C., RIMBOURG S., RUIZ-GAZEN A. and THOMAS-AGNAN C. (2013), "Fiches méthodologiques, Méthodes statistiques d'allocation spatiale : interpolation de données surfaciques", technical report.
- [29] 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<sup>o</sup> 99-26, Wisconsin, Madison.

- [30] WRIGHT J.K., 1936. "A method of mapping densities of population with Cape Cod as an example", *Geographical Review*, vol. 26, n<sup>o</sup> 1, pp. 103-110.
- [31] 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<sup>o</sup> 1, pp. 78-98.
- [32] YOO E., KYRIAKIDIS P.C., 2006, "Area-to-point kriging with inequality-type data", *Journal of geographical systems*, vol. 8, n<sup>o</sup> 4, pp. 357-390.
- [33] 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<sup>o</sup> 3, pp. 245-258.

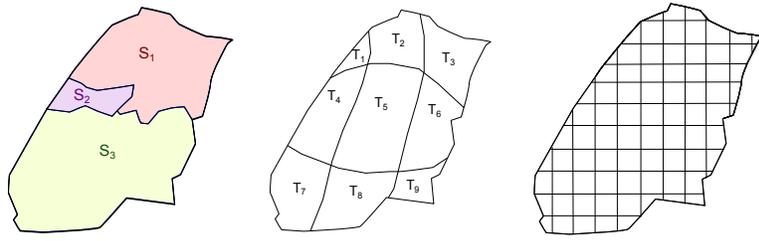


Figure 1: Source zones, target zones and grid target zones.

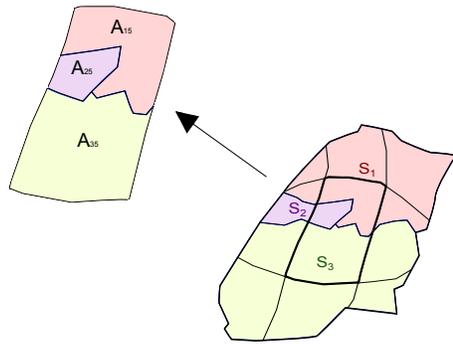


Figure 2: Intersection zones.

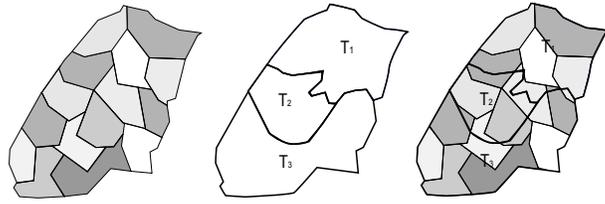


Figure 3: Aggregation case.

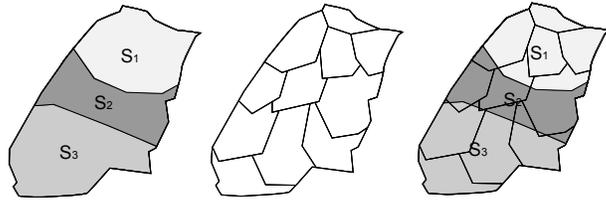


Figure 4: Disaggregation case.

196	204	156	136	113
163	135	112	116	113
144	108	91	112	107
143	95	112	96	85
131	111	88	91	92

122	91	67	60	58
75	58	36	32	31
71	32	21	15	11
51	31	22	8	9
56	30	9	10	5

117	115	73	57	41
80	56	30	31	34
61	25	22	13	17
63	37	15	11	8
62	28	11	8	6

Figure 5: Toy example. Data on cells.  $Y1$  (left),  $X$  (central),  $Y2$  (right)

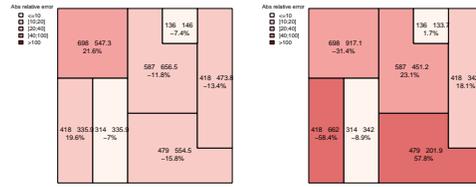


Figure 6: Toy example. Target variable  $Y_1$ : Areal weighting (left) and dasymetric with  $X$  (right)

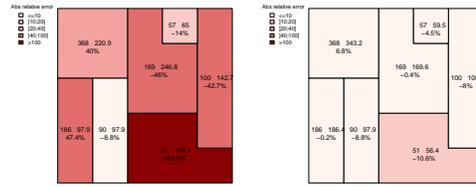


Figure 7: Toy example. Target variable  $Y_2$ : Areal weighting (left) and dasymetric with  $X$  (right)

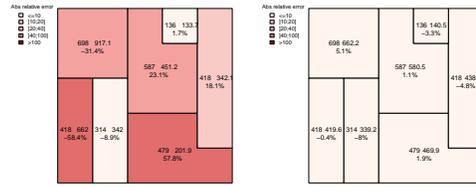


Figure 8: Toy example. Target variable  $Y_1$ : Dasymetric (left) and Regression (right)

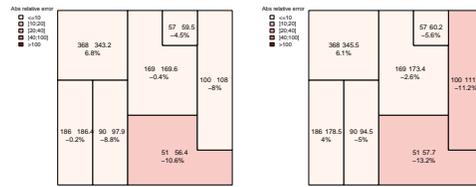


Figure 9: Toy example. Target variable  $Y_2$ : Dasymetric (left) and Regression (right)

Methods	Target variable Y		Auxiliary variable X				Control zones	Pycnophylactic property
	Nature	Additional assumptions	Dimension	Nature	Support			
Areal weighting	Extensive	Homogeneous on sources	none	none	none	none	yes	
	Intensive							
Dasymetric	Extensive	none	1	Extensive or intensive	intersection	none	yes	
	Intensive							
Dasymetric with control zone	Extensive	Homogeneous on controls	1	categorical	control	yes	no	
	Intensive	Homogeneous on controls						
Regression without auxiliary info	Extensive	Homogeneous	none	none	none	none	no	
	Intensive	on targets						
Regression with auxiliary info	Extensive	none	$\geq 1$	Extensive or intensive	target	none	no	
	Intensive	weight area	$\geq 1$	Extensive or intensive	target	none	no	
Point in polygon	Extensive	none	none	none	none	none	no	

Table 1: Summary of methods