

## AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur : ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite de ce travail expose à des poursuites pénales.

Contact : [portail-publi@ut-capitole.fr](mailto:portail-publi@ut-capitole.fr)

## LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10

Loi n°92-597 du 1<sup>er</sup> juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992

<http://www.cfcopies.com/V2/leg/leg-droi.php>

<http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm>



# THÈSE

En vue de l'obtention du

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 1 Capitole (UT1 Capitole)*

---

---

Présentée et soutenue le 15/06/2015 par :

**DO Van Huyen**

**Les Méthodes d'interpolation pour données sur zones**

---

---

### JURY

ANNE RUIZ-GAZEN	Professeur, Université Toulouse 1 Capitole	Président du Jury
ANNE VANHEMS	Professeur, Université de Toulouse et de Toulouse Business School	Membre du Jury
CHRISTINE THOMAS-AGNAN	Professeur, Université Toulouse 1 Capitole	Membre du Jury
EDITH GABRIEL	Maître de conférences, Université d'Avignon et des Pays de Vaucluse	Membre du Jury
FLORENT BONNEU	Maître de conférences, Université d'Avignon et des Pays de Vaucluse	Membre du Jury
GIUSEPPE ARBIA	Professor, Catholic University of Sacred Heart	Membre du Jury

---

### École doctorale et spécialité :

*MITT : Domaine Mathématiques : Mathématiques appliquées*

### Unité de Recherche :

*Groupe de Recherche en Économie Mathématique et Quantitative (GREMAQ)*

### Directeur(s) de Thèse :

*Christine THOMAS-AGNAN et Anne VANHEMS*

### Rapporteurs :

*Edith GABRIEL et Giuseppe ARBIA*

# Abstract

The combination of several socio-economic data bases originating from different administrative sources collected on several different partitions of a geographic zone of interest into administrative units induces the so called areal interpolation problem. This problem is that of allocating the data from a set of source spatial units to a set of target spatial units. At the European level for example, the EU directive 'INSPIRE', or INfrastructure for Spatial InfoRmation, encourages the states to provide socio-economic data on a common grid to facilitate economic studies across states. In the literature, there are three main types of such techniques: proportional weighting schemes, smoothing techniques and regression based interpolation. We propose a theoretical evaluation of these statistical techniques for the case of count related data. We find extensions of some of these methods to new cases : for example, we extend the ordinary dasymetric weighting method to the case of an intensive target variable  $Y$  and an extensive auxiliary quantitative variable  $X$  and we introduce a scaled version of the Poisson regression method which satisfies the pycnophylactic property. We present an empirical study on an American database as well as an R-package for implementing these methods.

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

## ACKNOWLEDGEMENTS

I would like to thank my advisors Christine Thomas and Anne Vahems for teaching me not only statistics but also the way to organize life of scientific women; to co-advisor Nguyen Tien Zung for lessons about many things (never related to mathematics); Do Duc Thai for introducing me to Zung and thanks to that, I met my wonderful supervisors.

I would like to thank my thesis committee members for all of their guidance through this process. Their valuable discussion, advices that they have given me make improvement not only for this thesis but also for my future research.

I would send my big thanks to Tibo and Kartika for making me come to the office more frequently because I knew they were there. With them I enjoyed France so much.

Je remercie Pascal, Maïti et Azalais pour m'avoir patiemment écouté parler en français pas toujours correct et pour m'avoir fait sentir comme à la maison.

I would like to thank all Vietnamese students, specially Chinh for the first helps in the very beginning days, and all others that if I numerate will last forever for giving me advices and encourage me in complicated administrative papers in France. I didn't shock when arriving France but got shocked when one by one of them finished their study and left France.

Je remercie Nga et Khue pour avoir été à mes côtés à Toulouse depuis l'avion qui nous a amené en France ensemble pour la première fois.

I would like to thank all professors, specially Michel Simioni, Anne Ruiz-Gazen, and all friends at my office for lessons that they have taught me. I would have hardly been able to learn so much about the world in anywhere but TSE.

Cảm ơn bố tôi vì luôn để tôi lựa chọn.

Cảm ơn Việt vì luôn tôn trọng những suy nghĩ khác biệt của hai chị em.

And finally, special thanks to Minh for making me feel so safe and so loved.



# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Interpolation problem . . . . .	7
1.2	Accuracy issue . . . . .	8
1.3	Notation . . . . .	10
<b>2</b>	<b>Review of areal interpolation methods</b>	<b>15</b>
2.1	Elementary methods . . . . .	17
2.1.1	Point in polygon . . . . .	17
2.1.2	Areal weighting interpolation . . . . .	17
2.2	Dasymetric weighting . . . . .	18
2.2.1	Ordinary dasymetric weighting . . . . .	19
2.2.2	Dasymetric weighting with control zones . . . . .	19
2.2.3	Two steps dasymetric weighting . . . . .	20
2.3	Regression techniques . . . . .	20
2.3.1	Regression without auxiliary information . . . . .	21
2.3.2	Regression with control zones . . . . .	21
2.3.3	Regression with auxiliary information at target level . . . . .	22
2.4	A short overview of more elaborate methods . . . . .	24
2.4.1	Other regression methods . . . . .	24
2.4.2	Smoothing techniques . . . . .	24
<b>3</b>	<b>Accuracy</b>	<b>27</b>
3.1	Count variable and model . . . . .	27
3.2	Relative accuracy of areal weighting and dasymetric: finite distance assessment . . . . .	30
3.2.1	General auxiliary information model . . . . .	30
3.2.2	Homogeneous model . . . . .	34
3.2.3	Piecewise homogeneous model . . . . .	35
3.3	Relative accuracy of the other methods: asymptotic assessment . . . . .	35
3.3.1	Estimators of the regression coefficients . . . . .	36
3.3.2	Predictors . . . . .	37
3.4	Simulated toy example . . . . .	39
<b>4</b>	<b>Application and R-package</b>	<b>45</b>
4.1	Application . . . . .	45
4.1.1	Data . . . . .	45
4.1.2	Results . . . . .	46
4.1.3	Spatial scale . . . . .	49
4.2	Package . . . . .	53

<b>5</b>	<b>Perspective and conclusion</b>	<b>59</b>
<b>6</b>	<b>Appendix</b>	<b>63</b>

# Chapter 1

## Introduction

### 1.1 Interpolation problem

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 databases 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 a new school or store should 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. A 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 1.3, we will see that there are three main types of such techniques in Chapter 2. 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. (1999), 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), Goodchild et al. (1993), Flowerdew and Green (1993) for the simplest ones. The third type comprises smoothing techniques which are described for example in Tobler (1979), Martin (1989), Bracken and Martin and Bracken (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. In this work we concentrate on the simple methods 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 2.1) 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 2.2.1, we extend the ordinary dasymetric weighting method to the case of an intensive target variable  $Y$  and an extensive auxiliary quantitative variable  $X$  and in section 2.2.2 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 in Chapter 3.

## 1.2 Accuracy issue

As mentioned above, one motivation of Chapter 2 is 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 et al. (1999), Mennis (2006), Fisher and Langford (1996), 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 (1993), 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 et al. (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 accuracy 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.

In the document, our objective is to analyze the accuracy of the simple interpolation methods with a methodological point of view.

Comparing the accuracy of the different methods is difficult because the relative accuracy depends on several factors: nature of the target variable, correlation between the target and auxiliary variables, shapes of zonal sets, relative size between the two zonal sets,... In order to derive theoretical results, we need to consider simplifying restrictions. For this reason, in this document, we first of all restrict attention to data obtained from counts (see Chapter 2): they are frequent in the literature and cover most of the cases in the socio-economic applications. We also restrict the comparison to the simplest classes of methods which are the dasymetric and the regression ones. At last, we make the assumption that target zones are nested within source zones. Indeed, this is not really a restriction since the intersections between sources and targets are always nested within sources and it is immediate to go from intersection level to target level by aggregating the predictions as we will see later. In Section 3.1, we define what we mean by data obtained from counts and we introduce a mathematical model adapted to this case. In order to illustrate the methods and check our theoretical results, we present a set of simulated data that we use later. Finally, in section 3.2, we compare the relative accuracy of areal weighting and dasymetric methods with finite distance results whereas in section 3.3, we compare the relative accuracy of dasymetric and Poisson regression methods with asymptotic methods. In both sections, we comment the results obtained on the toy examples presented in Section 3.1. All proofs are in the appendix.

The aim of Chapter 4 is to derive some empirical guidelines of application for several areal interpolation methods and confront the empirical evidence with some theoretical results of Chapter 3 using the demographic database ‘US census 2010’ (Almquist et al. (2010)) available in an R package. We concentrate on count related data and focus particularly on the following points. The data is presented in section 4.1.1.

In section 4.1.2 we give some directives for the selection of a good auxiliary variable when there is a choice. In section 4.1.2 we compare the accuracy of the regression methods. The regression methods developed for the extensive case are different from the ones adapted to the intensive case. Since it is easy to transform an extensive variable into a corresponding intensive one and reversely, we explore in section 4.1.2 the question of whether it is best to use the extensive-type or intensive-type regression method. Finally we investigate in section 4.1.3 the effect of the spatial scale.

An R-package including all functions for the implementation of the investigated methods is programmed and presented in Chapter 4. The package contains options adapted to realistic situations which are simplified in order to ease calculation of theoretical results.

### 1.3 Notation

Let us first introduce the terminology and notation used hereafter in the document. 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 1.1 illustrates these two partitions of the region of interest.

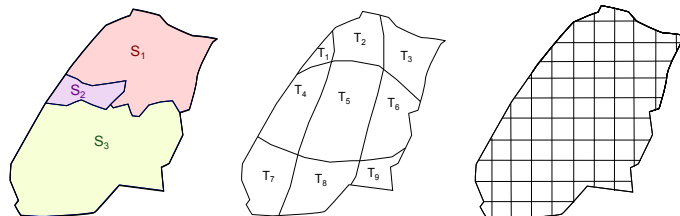


Figure 1.1: Source zones, target zones and grid target zones.

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 1.2 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  $A$ .

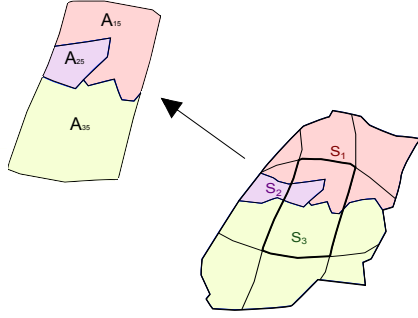


Figure 1.2: Intersection zones.

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

$$(1.1) \quad Y_\Omega = \frac{\sum_k P_{\Omega_k}}{|\Omega|} = \sum_k \frac{|\Omega_k|}{|\Omega|} \frac{P_{\Omega_k}}{|\Omega_k|} = \sum_{k=1}^p w_{\Omega_k} Y_{\Omega_k}.$$

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

$$(1.2) \quad w_{\Omega_k} = \frac{Z_{\Omega_k}}{Z_\Omega}.$$

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 (1.2), the variable

$$(1.3) \quad \tilde{Y}_A = \frac{Y_A}{Z_A}$$

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 (1.2), it can be transformed into an extensive variable by

$$(1.4) \quad \tilde{Y}_A = Z_A Y_A.$$

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 1.3 and 1.4 illustrate these cases. We will concentrate here on the disaggregation type.

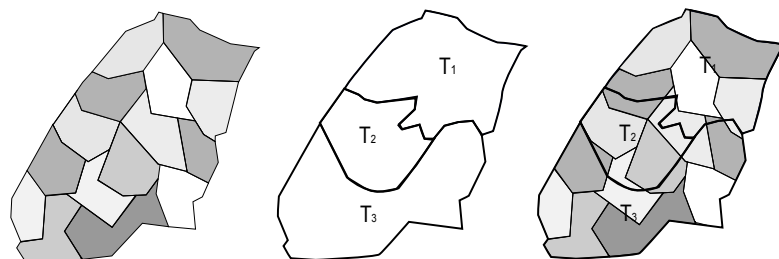


Figure 1.3: Aggregation case.

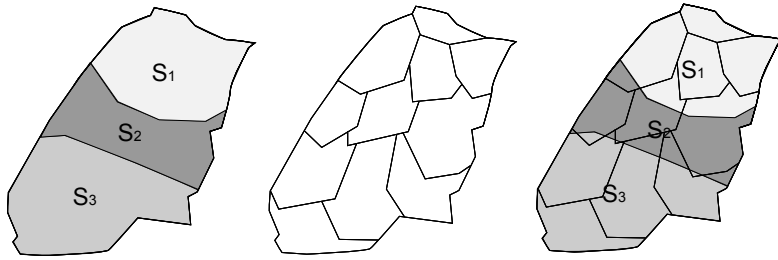


Figure 1.4: Disaggregation case.

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 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 (1.3) and (1.4).



## Chapter 2

# Review of areal interpolation methods

The areal interpolation problem has been studied widely in the literature because of the need for socio-economic analysis. In the chapter, we attempt to classify and formalize the most used and simple methods for the area-to-area case. Our aim is to give an overview of those methods which then helps us to evaluate theoretically their accuracy.

Let us start by introducing the toy example that will be used to demonstrate some properties. On Figure 2.1, we can see a square divided into 25 equal cells and three source regions made of unions of cells. The Figure presents 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 2.2 through 2.5 which compare some methods through the targets prediction errors. Precisely, the areal weighting and dasymetric methods for homogeneous and inhomogeneous cases are displayed in Figures 2.2 and 2.3, whereas Figures 2.4 and 2.5 draw the comparison between the dasymetric and regression methods.

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 2.1: Toy example. Data on cells.  $Y_1$  (left),  $X$  (central),  $Y_2$  (right)



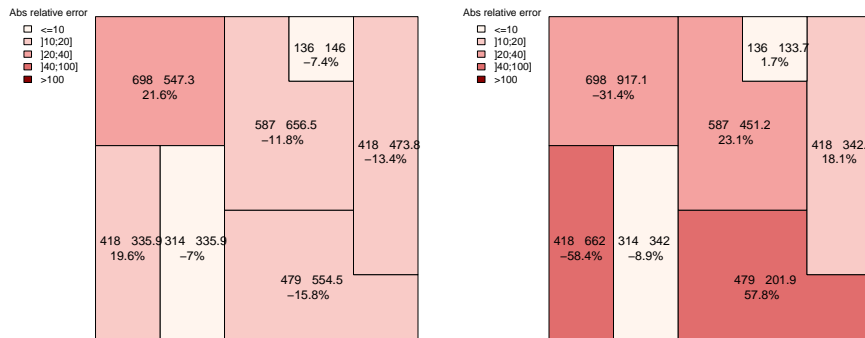


Figure 2.2: Toy example. Target variable  $Y_1$ : Areal weighting (left) and dasymetric with  $X$  (right)

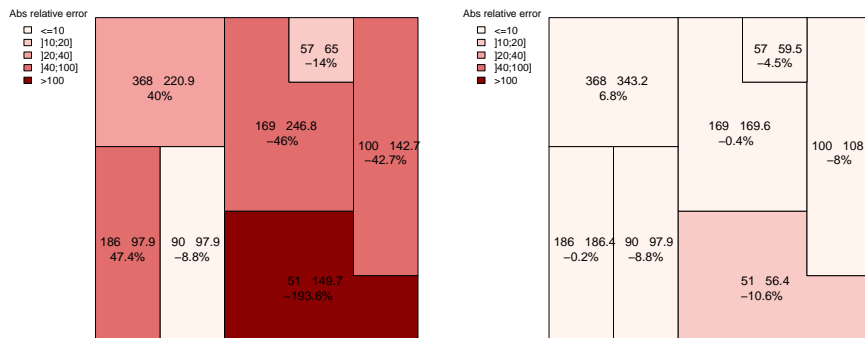


Figure 2.3: Toy example. Target variable  $Y_2$ : Areal weighting (left) and dasymetric with  $X$  (right)

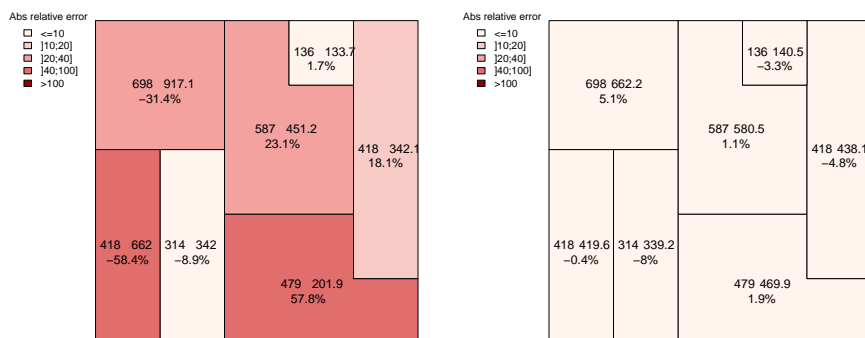


Figure 2.4: Toy example. Target variable  $Y_1$ : Dasymetric (left) and Regression (right)

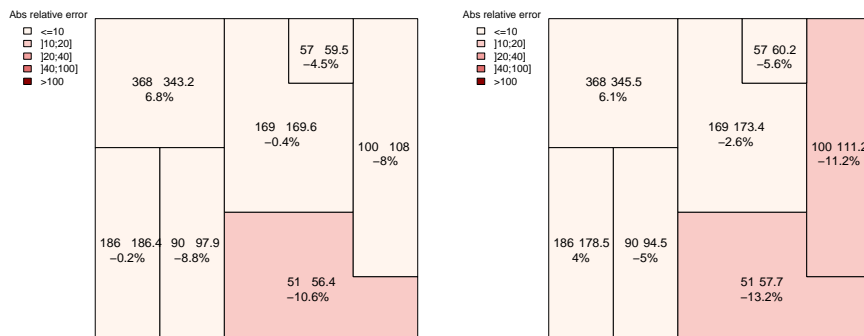


Figure 2.5: Toy example. Target variable  $Y_2$ : Dasymetric (left) and Regression (right)

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  $X$  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.

## 2.1 Elementary methods

### 2.1.1 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. (1999) report that it is the least accurate method. Moreover, it does not satisfy the pycnophylactic property.

### 2.1.2 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  $s \cap t \neq \emptyset$ ,

$$(2.1) \quad \hat{Y}_{s,t} = \frac{|A_{s,t}|}{|S_s|} Y_s.$$

After the combination step, this results in the following formula

$$(2.2) \quad \hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \frac{|A_{s,t}|}{|S_s|} Y_s.$$

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

$$(2.3) \quad \hat{Y}_{s,t} = Y_s.$$

After the combination step, this results in the following formula

$$(2.4) \quad \hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \frac{|A_{s,t}|}{|T_t|} Y_s.$$

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 2.3.1.

## 2.2 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 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.

### 2.2.1 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. (1999) 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 (2.1) and (2.2) by substituting  $X$  for the area:

$$(2.5) \quad \hat{Y}_{s,t} = \frac{X_{s,t}}{X_s} Y_s.$$

yielding after the combination step:

$$(2.6) \quad \hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \frac{X_{s,t}}{X_s} Y_s.$$

We propose to extend this method to the case of an intensive target variable with weights given by  $w_A = \frac{Z_A}{Z_Q}$  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 (1.3) (see the annex for a proof).

$$(2.7) \quad \hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \frac{X_{s,t}}{X_s} \frac{Z_s}{Z_t} Y_s.$$

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 beginning of this chapter. Figure 2.2 presents a comparison between the results of the areal weighting method and the dasymetric method for target variable  $Y_1$ . Figure 2.3 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.

### 2.2.2 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 (1.4) 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.

### 2.2.3 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.

## 2.3 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 2.3.1 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 2.3.2 is a regression version of the dasymetric weighting with control zones of section 2.2.2. The regression with auxiliary information at target level of section 2.3.3 can be compared to ordinary dasymetric weighting of section 2.2.1.

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)).

### 2.3.1 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 2.3.2.

### 2.3.2 Regression with control zones

Using control zones as in section 2.2.2, Goodchild et al. (1993) propose a two steps procedure where the first step is the technique of section 2.3.1 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 2.3.1. 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. (1997) apply rather a dasymmetric second step which they call “scaling” using the first step target variable prediction as an auxiliary variable, thus enforcing the pycnophylactic property. Reibel and Bufalino (2005) 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 these 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 pycnophylactic property. Then with an aggregation formula they get population estimates for any combination of grid cells, namely for target regions.

### 2.3.3 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 (1993), 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 (1.2) 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 (1993) (see paragraph 2.3.3), 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

$$(2.8) \quad \mathbb{E}(Y_{s,t} | Y_s) = \frac{\mu_{s,t}}{\sum_{t'} \mu_{s,t'}} Y_s$$

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

$$(2.9) \quad \mathbb{E}(Y_{s,t} | Y_s) = \mu_{s,t} + \frac{Cov(Y_{s,t}, Y_s)}{Var(Y_s)} (Y_s - \mu_s) = \mu_{s,t} + (Y_s - \mu_s)$$

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

$$(2.10) \quad \mu_s = \mathbb{E}(Y_s) = \mathbb{E} \left( \sum_t \frac{n_{s,t}}{n_s} Y_{s,t} \right) = \sum_t \frac{n_{s,t}}{n_s} \mu_{s,t}.$$

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 (2.10).

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 (1993) 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

$$(2.11) \quad \mu_S = W X_T \beta$$

between target means at source level and auxiliary information at target level. Using the data at the source level  $Y_S$  and equation (2.11), 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 2.4 (resp 2.5) 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 (1993) 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

$$(2.12) \quad \mu_C = X_C \beta$$

$$(2.13) \quad \mu_S = W X_C \beta$$

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 (2.13), 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  $\frac{n_{c,t}}{n_c}$ .

## 2.4 A short overview of more elaborate methods

### 2.4.1 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.

### 2.4.2 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 Martin and Bracken (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.

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	none	yes	
	Intensive								
Dasymeric	Extensive	none	1	Extensive or intensive	intersection	none	yes	yes	
	Intensive								
Dasymeric with control zone	Extensive	Homogeneous on controls	1	categorical	control	yes	no	no	
	Intensive	Homogeneous on controls							
Regression without auxiliary info	Extensive	Homogeneous	none	none	none	none	no	no	
	Intensive	on targets							
Regression with auxiliary info	Extensive	none	$>= 1$	Extensive or intensive	target	none	no	no	
	Intensive	weight area	$>= 1$	Extensive or intensive	target	none	no	no	
Point in polygon	Extensive	none	none	none	none	none	no	no	

Table 2.1: Summary of methods

## Chapter 3

# Accuracy

In the previous chapter, we classified areal interpolation methods into three groups: smoothing, dasymetric and regression based methods. On one hand, the review gives a clear picture about the simple available methods in literature, on the other hand, it prepares a mathematical base for a further step: the comparison of these methods. This chapter is aimed to study theoretically the accuracy of those methods for the count variable. We first describe the concept of count related data then introduce a model to study the case. The accuracy of methods will be analyzed in the last two sections with two different approaches: finite distance and asymptotic.

### 3.1 Count variable and model

Most of economic data collected at regional level result from aggregating point data and are only released in this aggregated form. Intuitively, let us say that a point data set is a set of a random number of random points in a given region of geographical space. The collection of corresponding numbers of such points in given subdivisions of this region is a count data set. For example with census data, a population count on a given zone is the number of inhabitants of the zone. This number is obtained from the knowledge of the addresses of these people. The collection of coordinates of such addresses is the underlying point data set. Examples of areal interpolation of population or subpopulation counts can be found for example in Goodchild et al. (1980); Langford (2007); Mennis and Hultgren (2006); Reibel and Agrawal (2007). Other types of counts are encountered frequently, for example number of housing units in Reibel and Bufalino (2005). Another frequent type of count related variable is the number of points per areal unit associated to a point data set: it is a density type variable. Examples of areal interpolation of population densities can be found in Yuan et al. (1997); Murakami and Tsutsumi (2011). An even more general type is when the variable is a ratio of counts such as number of doctors per patient. There is an easy one to one correspondence between a count variable and a density variable which allows to transform one type into the other so that any treatment of counts can be extended to densities and reversely. A count variable belongs to the family of extensive variables, which are variables whose value on a region is obtained by summing up its values on any partition into subregions (aggregation formula hereafter). A density variable belongs to the family of intensive variables, which are variables whose value on a region is obtained from values on any partition into subregions by a weighted sum (see Do et al. (2015) for more details). In the case of population density, the weights are given by the areas of the subregions of the partition. In the remainder of this paper, we will concentrate on pure count variables.

We introduce a model for an extensive count variable by assuming that there exists an underlying (unreleased) Poisson point process  $Z_Y$  (in the population example, the positions of the individuals of the population) and that the target variable  $Y$  on a subzone  $A$  is the number of points of  $Z_Y$  in  $A$ . For a partition  $\Omega_i, i = 1, 2, \dots, k$  of the region  $\Omega$ , the aggregation property of the extensive type is clearly satisfied

$$(3.1) \quad Y_\Omega = \sum_{i=1}^k Y_{\Omega_i}.$$

With the proposed Poisson point process assumption, for any zone  $A$ ,  $Y_A = \sum_i \mathbf{1}_A(Z_i)$  is a Poisson distributed random variable with mean  $\lambda_A = \int_A \lambda_{Z_Y}(s) ds$ , where  $\lambda_{Z_Y}$  is the intensity of the point process  $Z_Y$ .

This model implies that  $Y_A$  and  $Y_B$  are automatically independent for all disjoint couples of subregions  $A$  and  $B$  due to the Poisson process nature. We could use point process models with interaction effects while retaining the extensive property but we rather devote this article to this first case, keeping the interaction case for further developments.

As we will see in the next section, some methods we want to compare (dasymetric and univariate regression) make use of an auxiliary information. For the auxiliary variable  $X$  to be relevant, there must be some relationship between the target variable and the auxiliary variable. In many cases a categorical information is used such as land cover: Reibel and Agrawal (2007) and Yuan et al. (1997) use land cover type data on a 30 meters resolution grid, Mennis and Hultgren (2006) use 5 types of land cover obtained manually from aerial photography. Li et al. (2007) just use a binary information such as unpopulated versus populated zones. Reibel and Bufalino (2005) interpolate the 1990 census tract counts of people and housing using length of streets as auxiliary information. Mugglin and Carlin (1998) exploit population to interpolate the number of leukemia cases. The use of a continuous auxiliary information can also be found: Murakami and Tsutsumi (2011) utilize distance and land price to predict population density. In the rest of the paper, we concentrate on a single extensive auxiliary variable  $X$  that is also a count in order to be able to consider the accuracy of all methods simultaneously (more details at the beginning of section 3.3). Therefore it corresponds to another underlying point process  $Z_X$  with intensity  $\lambda_{Z_X}$ .

The auxiliary variable  $X$ , has to be known at intersection level in the case of dasymetric and at the target level in the case of regression. We need to write a formal relationship between our target variable and the auxiliary information. The model we propose assumes that the following relationship holds between  $Y$  given  $X$ : at the level of any subset  $A$  of the region, the conditional distribution of  $Y_A$  given  $X_A = x_A$  is assumed to be

$$(3.2) \quad Y_A | X_A = x_A \sim \mathcal{P}(\alpha|A| + \beta x_A)$$

and this implies that the following relationship holds between the two underlying point processes intensity functions

$$(3.3) \quad \lambda_{Z_Y}(u) = \alpha + \beta \lambda_{Z_X}(u),$$

where  $u$  is any location in  $\Omega$ .

This relationship will be used at target level  $A = T$  and at source level  $A = S$ . This model in its general form will be called auxiliary information model (AIM). In this model, the intensity of  $Z_Y$  is driven by two effects: the effect of the auxiliary variable  $X$  and the effect of the area of the zone. If we look at target level, the target variable is Poisson distributed with a mean

comprising two parts  $\mathbb{E}(Y_T) = \alpha|T| + \beta x_T$ : the first part  $\alpha|T|$  reflects the impact of the area of the zone  $T$ , whereas  $\beta x_T$  is the impact of the auxiliary variable. The linearity of the expected value of  $Y$  with respect to the area and to the auxiliary information is not canonical in a Poisson regression model for counts but in our case it derives naturally from (3.3).

In sections 3.2.2 and 3.2.3, we introduce two sub-models of model (3.2) depending on the intensity function  $\lambda_{Z_Y}$ . We consider the case of a constant intensity (homogeneous model) and the case of a piecewise constant intensity (piecewise homogeneous model).

Concerning the regression based methods, there are several types of regression based methods also involving auxiliary information that we presented in the chapter 2. Given the nature of the target variable in our model (3.2), we concentrate on the Poisson regression presented in Flowerdew et al. (1991) for the purpose of predicting population (which is an extensive variable) with categorical auxiliary information. Based on model (3.2), a Poisson regression with identity link is performed at source level yielding estimators  $\hat{\alpha}, \hat{\beta}$  for the parameters  $\alpha$  and  $\beta$ .

The prediction of the target variable at intersection level is then obtained by

$$(3.4) \quad \hat{Y}_{st}^{REG} = \hat{\alpha}|A_{st}| + \hat{\beta}X_{st}$$

and the final step aggregates intersections predictions at target levels. The regression based methods can be considered as more powerful than the dasymetric methods in the sense that they can incorporate multivariate auxiliary information and that the knowledge of auxiliary information is only needed at target level and not at intersection level. However, the purpose of this paper being to compare the accuracy of dasymetric methods and Poisson regression methods from a methodological point of view and for the case of extensive count data, we therefore concentrate on the unidimensional auxiliary count variable case.

## Accuracy criterion

The accuracy assessment necessitates the choice of a prediction error criterion and of a geographic level. In this framework, examples of criteria are root mean square error or mean square error (Sadahiro (1999), Reibel and Bufalino (2005),...) at regional level (that is the union of all sources), or relative absolute error at target level (Langford, 2007). We denote by MET a generic method of prediction and let  $MET$  be  $DAW$  for the areal weighting method,  $DAX$  for the general dasymetric method,  $REG$  for the Poisson regression method and  $ScR$  for the scaled regression method which will be presented later in section 3.3. We recall that we assume all target zones are nested within source zones.

In section 3.2, we use mean square error at source level to compare the areal weighting and dasymetric methods. For method MET, the source level error is then computed as follows

$$(3.5) \quad \text{Er}_S^{MET} = \sum_{t \subset S} \text{Er}_t^{MET} = \sum_{t \subset S} \mathbb{E}(\hat{Y}_t^{MET} - Y_t)^2$$

and the overall regional error is

$$(3.6) \quad \text{Er}^{MET} = \sum_S \sum_{t \in S} \mathbb{E}(\hat{Y}_t^{MET} - Y_t)^2$$

In section 3.3, we use mean square error at target level

$$(3.7) \quad \text{Er}_t^{MET} = \mathbb{E}(\hat{Y}_t^{MET} - Y_t)^2$$

to compare the dasymetric and Poisson regression methods.

In general, we will also use the relative error criterion defined as

$$(3.8) \quad \text{Re}_S^{MET} = \frac{\sqrt{\text{Er}_S^{MET}}}{\mathbb{E}(Y_S)}$$

where  $\text{Re}_S^{MET}$  is the relative error of method  $MET$  at source level for source  $S$  with method  $MET$ .

## 3.2 Relative accuracy of areal weighting and dasymetric: finite distance assessment

Let us briefly summarize the findings of the assessments found in the literature for the comparison of general dasymetric and areal weighting. For empirical assessments, several authors report that the dasymetric method improves upon areal weighting. Depending on the context, the improvement varies: Langford (2007) reports improvements of 54%, 57%, and 59% better depending on the auxiliary information used; Reibel and Bufalino (2005) reports improvements of 71.26% and 20.08% with street length auxiliary information for the two target variables: housing units and total population. For theoretical assessments, (Sadahiro, 1999, 2000) compares the areal weighting interpolation and the point-in-polygon method with a theoretical model. We did not mention yet the point-in-polygon method because it is a very elementary one consisting in allocating a source value to the target which contains its centroid. Using a stochastic model, he finds that the factors that impact the accuracy of the methods are the size and shape of target and source zones, the properties of underlying points.

In this section, we prove some theoretical properties in subsection 3.2.1, with two particular cases in 3.2.2 and 3.2.3.

Since targets are nested within sources, the predictors of the two methods depend only on the source that contains the concerned target zone. For that reason, we focus on studying one source zone denoted by  $S$ . For a target  $T$  in  $S$ , the two predictors are as follows

$$(3.9) \quad \hat{Y}_t^{DAW} = \frac{|T|}{|S|} Y_s$$

and

$$(3.10) \quad \hat{Y}_t^{DAX} = \frac{x_T}{x_S} Y_s.$$

### 3.2.1 General auxiliary information model

Lemma 3.2.1 gives the expression of the prediction bias and variance in model AIM for areal weighting interpolation and dasymetric interpolation at target level.

**Lemma 3.2.1.** *In model AIM, the prediction biases and variances at target level of the areal*

weighting interpolation and dasymetric methods are given by

$$(3.11) \quad \mathbb{E}(\hat{Y}_T^{DAW} - Y_T) = \beta x_S \left( \frac{|T|}{|S|} - \frac{x_T}{x_S} \right)$$

$$(3.12) \quad \mathbb{E}(\hat{Y}_T^{DAX} - Y_T) = \alpha |S| \left( \frac{x_T}{x_S} - \frac{|T|}{|S|} \right)$$

$$(3.13) \quad \text{Var}(\hat{Y}_T^{DAW} - Y_T) = \beta x_S \left( \frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_T \left( 1 - \frac{x_T}{x_S} \right) + \alpha |T| \left( 1 - \frac{|T|}{|S|} \right)$$

$$(3.14) \quad \text{Var}(\hat{Y}_T^{DAX} - Y_T) = \alpha |S| \left( \frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_T \left( 1 - \frac{x_T}{x_S} \right) + \alpha |T| \left( 1 - \frac{|T|}{|S|} \right).$$

First note that the two biases have opposite signs, in other words, if the areal weighting interpolation method underestimates then the dasymetric method overestimates and vice versa. This fact can be interpreted as follows: while the true intensity comprises two effects, these methods treat only one of them which causes the contrast. Although the signs of biases are opposite, their absolute values are both proportional to  $\left| \frac{|T|}{|S|} - \frac{x_T}{x_S} \right|$  which measures the divergence between the share of the auxiliary information in target  $T$  with respect to  $S$  and the share of the area of  $T$  with respect to  $S$ . This divergence is also proportional to  $\frac{x_S}{|S|} - \frac{x_T}{|T|}$  and hence can be viewed as a distance to proportionality between area and auxiliary information. The bias of the areal interpolation method with its assumption of homogeneity is independent in the areal effect  $\alpha|S|$  but is proportional to the ignored auxiliary information effect, and reversely the dasymetric method which focuses on the effect of the auxiliary information gets rid of the  $\beta x_S$  in its bias but is proportional to the ignored areal effect. We will build on this to propose a new method in the next section.

The two variances have a common part  $\beta x_T \left( 1 - \frac{x_T}{x_S} \right) + \alpha |T| \left( 1 - \frac{|T|}{|S|} \right)$  which we can interpret as the loss of information when transferring data from a large source zone to a smaller target zone. For the remaining part, the same explanations as for the bias stands. Both variances have a parabola shape with respect to  $x_T$  (respectively to  $|T|$ ) with a maximum at  $x_T = \frac{1}{2}x_S$ , (resp.  $|T| = \frac{1}{2}|S|$ ): we can say loosely that the variances are maximum when the target zone is around a haft of the source. They vanish when the target zone is either empty or coincide with the source which makes sense. The reallocation to a larger target intuitively decreases the difficulty of the disaggregation problem except that the error also depends on the expected number of points so we should turn attention to relative error. If one divides the variances by the square of the expected number of points in the target zone  $\mathbb{E}(Y_T)$ , we can see that the relative error will tend to zero as  $\mathbb{E}(Y_T)$  tends to infinity.

Since the dasymetric method is pycnophylactic, the bias at source level is zero. Lemma 3.2.2 reports the expression of the prediction variances in model AIM for areal weighting interpolation and dasymetric interpolation at source level.

**Lemma 3.2.2.** *In model AIM, the variances of the areal weighting and dasymetric methods at the source level are*

$$(3.15) \quad \text{Var}(\hat{Y}_S^{DAW} - Y_S) = \beta x_S \sum_T \left( \frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_S \left( 1 - \sum_T \frac{x_T^2}{x_S^2} \right) + \alpha |S| \left( 1 - \sum_T \frac{|T|^2}{|S|^2} \right)$$

$$(3.16) \quad \text{Var}(\hat{Y}_S^{DAX} - Y_S) = \alpha |S| \sum_T \left( \frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_S \left( 1 - \sum_T \frac{x_T^2}{x_S^2} \right) + \alpha |S| \left( 1 - \sum_T \frac{|T|^2}{|S|^2} \right).$$



To get an insight at the impact of the number  $n_T$  of the target zones, we consider the particular case where all targets have the same size. In this case,  $\frac{|T|}{|S|} = \frac{1}{n_T}$  for any  $T$ , and we get

$$\begin{aligned} \text{Var}(\hat{Y}_S^{DAW} - Y_S) &= \left(1 - \frac{1}{n_T}\right)(\alpha|S| + \beta x_S) \\ \text{Var}(\hat{Y}_S^{DAX} - Y_S) &= \left(1 - \frac{1}{n_T}\right)(\alpha|S| + \beta x_S) + \sum_T \left(\frac{x_T^2}{x_S^2} - \frac{1}{n_T}\right)(\alpha|S| - \beta x_S). \end{aligned}$$

It is obvious that the larger the number of the target zones, the larger the variances, which agrees with our conclusion concerning the size of targets. Indeed, when the area of the target zones gets smaller, the error on each target decreases but the total error at the source level gets larger due to the effect of the number of the targets.

We are now ready to compute the mean square error difference between the two methods. We introduce the following quantities which quantify, at the geographical level of a subregion  $A$ , a relative contribution of each effect to the overall mean (respectively to the intensity):

$$I_A(X) = \frac{\beta x_A}{\alpha|A| + \beta x_A}.$$

$I_A(X)$  is the relative contribution of variable  $X$  and similarly  $I_A(|\cdot|) = \frac{\alpha|A|}{\alpha|A| + \beta x_A}$  is the relative contribution of the areal effect.

The imbalance between the two effects is measured by the difference

$$\Delta_A = I_A(|\cdot|) - I_A(X) = \frac{\alpha|A| - \beta x_A}{\mathbb{E}(Y_A)}.$$

This quantity ranges between  $-1$  when there is a pure  $X$  effect and  $1$  when there is a pure areal effect with a value of zero when the two effects are of equal size.

We can derive from lemmas 3.2.1 and 3.2.2 the expression of the absolute and relative errors of the two methods at source level as a function of the relative contributions.

**Theorem 3.2.3.** *In model AIM, the errors and relative errors of the areal weighting and dasy-metric methods at the source level are*

$$(3.17) \quad Er_S^{DAW} = I_S(X)^2 \mathbb{E}(Y_S)^2 D + I_S(X) \mathbb{E}(Y_S)(D + B) + I_S(|\cdot|) \mathbb{E}(Y_S) C,$$

$$(3.18) \quad Er_S^{DAX} = I_S(|\cdot|)^2 \mathbb{E}(Y_S)^2 D + I_S(|\cdot|) \mathbb{E}(Y_S)(D + C) + I_S(X) \mathbb{E}(Y_S) B,$$

$$(3.19) \quad (Re_S^{DAW})^2 = I_S(X)^2 D + \frac{1}{\mathbb{E}(Y_S)} [I_S(X)(D + B - C) + C],$$

$$(3.20) \quad (Re_S^{DAX})^2 = I_S(|\cdot|)^2 D + \frac{1}{\mathbb{E}(Y_S)} [I_S(|\cdot|)(D - B + C) + B],$$

where  $D = \sum_T \left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2$ ,  $B = 1 - \sum_T \frac{x_T^2}{x_S^2}$ ,  $C = 1 - \sum_T \frac{|T|^2}{|S|^2}$  are positive.

Note that  $B, C$  and  $D$  only depend on the geometry of the problem and the auxiliary information, whereas the relative contribution terms and  $\mathbb{E}(Y_S)$  depend on the coefficients  $\alpha$  and  $\beta$ . It is interesting to mention the symmetry between the two methods which stands clearly in these formulas when we exchange the two contributions terms. One can derive from this theorem the difference between the relative errors of the two methods

$$(3.21) \quad (\text{Re}_S^{DAW})^2 - (\text{Re}_S^{DAX})^2 = -D\Delta_S \left(1 + \frac{1}{\mathbb{E}(Y_S)}\right)$$

which turns out to be clearly proportional to the imbalance term  $\Delta_S$ . Similarly, one can approximate the ratio of the two relative errors when  $\mathbb{E}(Y_S)$  is large and  $D$  is not too small on the target  $A = T$  and on the source  $A = S$  by

$$(3.22) \quad \frac{\text{Re}_A^{DAW}}{\text{Re}_A^{DAX}} \approx \frac{I_A(X)}{I_A(|\cdot|)}.$$

This ratio roughly ranges from 0 to  $+\infty$  at the extreme cases of a pure  $X$  or areal effect showing that one can outperform the other by a large amount.

Let us now turn attention to the difference between the two errors.

**Theorem 3.2.4.** *The difference between the errors of areal weighting and dasymetric methods on a target zone  $T$  is*

$$\text{Er}_T^{DAW} - \text{Er}_T^{DAX} = \left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 \Delta_S \mathbb{E}(Y_S)(\mathbb{E}(Y_S) + 1).$$

The important conclusion of this result is that the sign of the difference in error agrees with the sign of  $\Delta_S$ , i.e. the sign of  $(\alpha|S| - \beta x_S)$ . Moreover, for  $\Delta_S < 0$ , as the effect of the auxiliary information  $I_S(X)$  gets stronger, the dasymetric method gets better and the difference between the two methods larger.

This computation result leads to a very interesting consequence: if one of two effect dominates on a given source, the related method wins on all target zones belonging to this source. It also shows that two methods will have the same accuracy if the two effects are balanced or the auxiliary variable is homogeneous.

The normalized difference between the two effects  $\Delta_S$  clearly determines which method is the best.

At this point, it seems natural to look for a linear combination of these two predictors

$$(3.23) \quad \hat{Y}_T^C(w) = w\hat{Y}_T^{DAW} + (1-w)\hat{Y}_T^{DAX},$$

which would combine their good properties. It turns out that in the class of linear combinations of areal weighting and dasymetric predictors, the best predictor is given by the following theorem

**Theorem 3.2.5.** *In model AIM, the best predictor in the sense of minimizing (with respect to the weight  $w$ ) the errors on any target zone  $T$  in the class (3.23) is*

$$(3.24) \quad \hat{Y}_T^C = \hat{Y}_T^C(w^*) = \frac{\alpha|T| + \beta x_T}{\alpha|S| + \beta x_S} Y_S$$

for  $w^* = \frac{\alpha|T|}{\alpha|S| + \beta x_S}$ . Its error and relative error are respectively given by

$$(3.25) \quad \text{Er}_T^C = \frac{\mu_T(\mu_S - \mu_T)}{\mu_S}$$

$$(3.26) \quad (\text{Re}_S^C)^2 = \frac{1}{4\mathbb{E}(Y_S)} [\Delta_S^2 D + 2\Delta_S(C - B) + D + 2B + 2C],$$

where  $\mu_A = \mathbb{E}(Y_A)$ . Moreover, this predictor coincides with the best linear unbiased predictor in model AIM.

Because  $\frac{\mu_T(\mu_S - \mu_T)}{\mu_S} = \text{Var}(\hat{Y}_T^{DAX} - Y_T) - \mu_S(\frac{x_T}{x_S} - \frac{\mu_T}{\mu_S})^2 = \text{Var}(\hat{Y}_T^{DAW} - Y_T) - \mu_S(\frac{|T|}{|S|} - \frac{\mu_T}{\mu_S})^2$ , the prediction error of the best predictor is smaller than the variances of the other two methods and the distance is all the more important that the auxiliary information is further from homogeneity. Of course, the oracle predictor  $\hat{Y}_T^C$  is not feasible since it depends on the unknown coefficients  $\alpha$  and  $\beta$  of model AIM but we will use it as a benchmark tool on the one hand and we will relate it later on to our new regression predictor. If we look at the error at the level of source  $S$ , we have that  $\text{Er}_S^C = \mu_S - \sum_T \frac{\mu_T^2}{\mu_S} \leq \mu_S - \frac{\mu_S}{n_T(S)}$ , where  $n_T(S)$  is the number of targets in source  $S$ , and hence this predictor's accuracy is worse when all targets have the same expected number of points  $\frac{\mu_S}{n_T(S)}$ . It is interesting to note that the relative error (at source level  $S$ ) of the best predictor tends to zero as the expected number of points in the source  $S$  tends to infinity, which was not the case for the dasymetric methods. For a fixed expected number of points in a given source  $S$ , we can easily find the value of the imbalance  $\Delta_S$  which minimizes

the relative error of  $\hat{Y}_T^C$   $\Delta^* = \frac{B - C}{D} = \frac{\sum_T (\frac{|T|}{|S|})^2 - (\frac{x_T}{x_S})^2}{\sum_T (\frac{|T|}{|S|} - \frac{x_T}{x_S})^2}$  and thus derive a lower bound for

the relative error for a given geometry. Some of these results are illustrated in Section 3.4. Because intuitively, it is natural to think that areal weighting should be outperformed by dasymetric when the underlying process is inhomogeneous, we consider the two cases of homogeneous and piecewise homogeneous submodels.

### 3.2.2 Homogeneous model

Areal weighting interpolation is a simple and natural rule which is based on the assumption that the target variable is homogeneous at source level. Indeed in model AIM, it is equivalent to assume that the point process is homogeneous and its intensity is therefore constant (equal to  $\alpha > 0$ ) leading to:

$$Y_A \sim \mathcal{P}(\alpha|A).$$

Substituting  $\beta = 0$  in (3.11), (3.13), (3.15) we get the bias, variance and error in this case:

$$\begin{aligned} \mathbb{E}(\hat{Y}_T^{DAW} - Y_T) &= 0 \\ \text{Er}_T^{DAW} &= \text{Var}(\hat{Y}_T^{DAW} - Y_T) = \alpha|T|(1 - \frac{|T|}{|S|}) \\ \text{Er}_S^{DAW} &= \text{Var}_S^{DAW} = \alpha|S|(1 - \sum_T \frac{|T|^2}{|S|^2}). \end{aligned}$$

Since  $\frac{1}{n_T} \leq \sum_T \frac{|T|^2}{|S|^2} \leq 1$ , the error at source level is maximum when all target zones have the same size, and minimal when there is a unique target which coincides with the source.

Substituting  $\beta = 0$  in (3.24) leads to the conclusion that the best linear unbiased predictor in the homogeneous AIM model is given by the areal weighting method which is a natural result. Let us now turn attention to a very simple non homogeneous model to illustrate the intuitive fact that the areal weighting interpolation method is not the best choice in a non homogeneous situation.

### 3.2.3 Piecewise homogeneous model

Suppose the source zone  $S$  comprises two homogeneous subzones  $C_1$  and  $C_2$  called control zones with intensities  $\alpha_1$  and  $\alpha_2$  respectively. In this case, we get

$$Y_A \sim \mathcal{P}(\alpha_*|A|),$$

where  $A \subset C_*$  with  $* = 1, 2$ . For simplification reasons, we assume the target zones to be nested within the control zones. The results of lemmas 3.2.1 and 3.2.2 give in this case

$$\mathbb{E}(\hat{Y}_T^{DAW} - Y_T)_{T:T \subset C_1} = \frac{|T|}{|S|}(\alpha_2 - \alpha_1)|C_2|$$

$$\mathbb{E}(\hat{Y}_T^{DAW} - Y_T)_{T:T \subset C_2} = \frac{|T|}{|S|}(\alpha_1 - \alpha_2)|C_1|$$

$$Var(\hat{Y}_S^{DAW} - Y_S) = \alpha_1|C_1|(1 - \sum_{T:T \subset C_1} \frac{|T|^2}{|S|^2}) + \alpha_2|C_2|(1 - \sum_{T:T \subset C_2} \frac{|T|^2}{|S|^2})$$

$$Er_S^{DAW} = Var(\hat{Y}_S^{DAW} - Y_S) + \sum_{T:T \subset C_1} \frac{|T|^2}{|S|^2}(\alpha_2 - \alpha_1)^2|C_2|^2 + \sum_{T:T \subset C_2} \frac{|T|^2}{|S|^2}(\alpha_2 - \alpha_1)^2|C_1|^2$$

The variance has a similar structure to the one of the homogeneous model. The bias clearly shows that the difference between the two intensities of the subzones will drive the size of the error.

## 3.3 Relative accuracy of the other methods: asymptotic assessment

Let us now try to extend the comparison to the Poisson regression methods. This cannot be done anymore by finite distance computations and so we introduce an asymptotic framework. Model (3.2) yields at source level

$$(3.27) \quad Y_s \sim \mathcal{P}(\alpha|S_s| + \beta x_s),$$

where  $x_s = \sum_{t:t \cap s \neq \emptyset} x_{st}$ . Besides the Poisson regression predictor defined by (3.4), inspired by Theorem 3.2.5, we propose a new predictor called scaled Poisson regression predictor defined as follows

$$(3.28) \quad \hat{Y}_{st}^{ScR} = \frac{\hat{\alpha}|A_{st}| + \hat{\beta}x_{st}}{\hat{\alpha}|S_s| + \hat{\beta}x_s} Y_S,$$

where  $\hat{\alpha}$  and  $\hat{\beta}$  are the estimators of  $\alpha$  and  $\beta$  obtained through the Poisson regression at source level. Note that if model (3.3) contains only one of the two effects (that of  $X$  for example), then it is easy to see that the predictor of the scaled regression method coincides with the dasymmetric method (corresponding to  $X$ ):

$$\hat{Y}_T^{ScR} = \frac{\hat{\beta}x_T}{\hat{\beta}x_S} Y_S = \hat{Y}_T^{DAX}.$$

In section 3.3.1, we establish the asymptotic properties of the estimators  $\hat{\alpha}$  and  $\hat{\beta}$  and these results will enable us to compare the predictors in section 3.3.2. Section 3.4 illustrates these results on a toy example.

### 3.3.1 Estimators of the regression coefficients

In this section, we adapt proofs from Fahrmeir and Kaufmann (1985) to establish the consistency and asymptotic normality of the estimators  $\hat{\alpha}, \hat{\beta}$ . We first need to describe an asymptotic framework. To be realistic, we assume that the whole region  $\Omega$  is fixed and that the number of source zones  $n_S$  (hereafter denoted by  $n$ ) increases to infinity. In this section, the source zones will be denoted by  $S_{n,i} : i = 1, 2, \dots, n$  and  $\Omega = \cup_i S_{n,i}$ . Because of the extensive property of  $X$ , we also assume a similar property of  $X_{n,i}$ : the total auxiliary information on the region  $\Omega$  remains constant  $x_\Omega = \sum_i x_{n,i}$ . In order to get a consistent regression however we need the amount of information at source level to increase and we thus assume that the intensity of  $Y$  increases with a rate  $k_n \rightarrow \infty$  so that

$$Y_A \sim \mathcal{P}(\alpha|\widetilde{A}| + \beta\tilde{x}_A),$$

where  $|\widetilde{A}| = k_n|A|, \tilde{x}_A = k_n x_A$ .

Let  $\gamma = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, Z_A = \begin{pmatrix} |A| \\ x_A \end{pmatrix}, Z_{n,i} = \begin{pmatrix} |S_{n,i}| \\ x_{n,i} \end{pmatrix}$ . With these notations we have  $\mu_A = \gamma' Z_A, \tilde{Z}_A = \begin{pmatrix} |\widetilde{A}| \\ \tilde{x}_A \end{pmatrix} = k_n Z_A$  and  $Y_A \sim \mathcal{P}(k_n \mu_A)$ . The true value of the parameter  $\gamma$  will be denoted by  $\gamma_o = \begin{pmatrix} \alpha_o \\ \beta_o \end{pmatrix}$ .

The log likelihood function  $l_n(\gamma)$ , the score function  $s_n(\gamma)$  and the information matrix  $F_n(\gamma)$  are then given by

$$l_n(\gamma) = \sum_{i=1}^n y_{n,i} \ln(\gamma' \tilde{Z}_{n,i}) - \gamma' \tilde{Z}_{n,i} - \ln(y_{n,i}!)$$

$$s_n(\gamma) = \frac{\partial l_n(\gamma)}{\partial \gamma} = \sum_{i=1}^n \frac{\tilde{Z}_{n,i}}{\gamma' \tilde{Z}_{n,i}} y_{n,i} - \tilde{Z}_{n,i}$$

$$F_n(\gamma) = \text{Cov}_\gamma(s_n(\gamma)) = \sum_{i=1}^n \frac{\tilde{Z}_{n,i} \tilde{Z}'_{n,i}}{\gamma' \tilde{Z}_{n,i}}$$

Differentiation of the score yields

$$H_n(\gamma) = -\frac{\partial s_n(\gamma)}{\partial \gamma} = \sum_{i=1}^n \frac{\tilde{Z}_{n,i} \tilde{Z}'_{n,i}}{(\gamma' \tilde{Z}_{n,i})^2} y_{n,i}$$

Our asymptotic framework differs from that of Fahrmeir and Kaufmann (1985) in the sense that at each step they have one new observation whereas in our case at each step all observations are new and we have one more than at the previous step. For this reason, we modify slightly their conditions and assume that

(C1)  $\{\tilde{Z}_{n,i}\} \subset \mathcal{Z} \forall n, i$  where  $\mathcal{Z}$  is a compact set.

(C2)  $\lambda_{\min}(\sum_i \tilde{Z}'_{n,i} \tilde{Z}_{n,i}) \rightarrow \infty$  as  $n \rightarrow \infty$  where  $\lambda_{\min}(W)$  denotes the minimum eigenvalue of the matrix  $W$ .

Condition (C1) is satisfied if there exists two positive numbers  $c_1, c_2$  (note that  $\|\tilde{Z}_{n,i}\| \neq 0$ ) s.t.

$$(3.29) \quad c_1 < \|\tilde{Z}_{n,i}\| < c_2.$$

In that case, the number of source zones increases with the rate of growth of the intensity at a similar rate and the number of points in one source zone is quite stable during the change process.

Under these conditions, we get the following asymptotic behavior for the Poisson regression coefficients.

**Theorem 3.3.1.** *Under conditions (C1) and (C2), the following statements holds for the Poisson regression estimator  $\hat{\gamma}_n$  of  $\gamma$*

(i)  $\hat{\gamma}_n \rightarrow_p \gamma_o$  (weak consistency)

(ii)  $F_n^{1/2}(\hat{\gamma}_n - \gamma_o) \rightarrow_d \mathcal{N}(0, \mathbf{I})$ . (asymptotic normality)

In the next section, we use these results to study the asymptotic behavior of the predictors.

### 3.3.2 Predictors

In this section, we consider the asymptotic properties of the following two predictors: the regression predictor (3.4) and the scaled regression predictor (3.28). We prove that the scaled regression predictor  $\hat{Y}_T^{ScR}$  is asymptotically as accurate as the unfeasible oracle predictor  $\hat{Y}_T^C$ . We also compare these two methods with areal weighting interpolation and dasymetric interpolation.

The first proposition is concerned with the pycnophylactic property, which is of interest in the areal interpolation literature.

**Proposition 3.3.2.** *The scaled Poisson regression predictor satisfies the pycnophylactic property at source level. The ordinary Poisson regression predictor is pycnophylactic at region level and asymptotically pycnophylactic at source level.*

We now turn attention to the asymptotic behavior of the prediction error for the ordinary Poisson regression predictor.

**Theorem 3.3.3.** *The asymptotic normality of the prediction error of the Poisson regression predictor at source level is given by*

$$\frac{\hat{Y}_{ni}^{REG} - Y_{ni}}{\sqrt{\gamma'_o \tilde{Z}_{n,i}}} \rightarrow_d \mathcal{N}(0, 1).$$

If we also assume a lower bound for  $\tilde{Z}_T$ , the following similar result at the target level holds

$$\frac{\hat{Y}_T^{REG} - Y_T}{\sqrt{\gamma'_o \tilde{Z}_T}} \rightarrow_d \mathcal{N}(0, 1).$$

The next result is about the quadratic prediction error and relative prediction error of the Poisson regression predictor.

**Theorem 3.3.4.** *For any  $\eta > 0$ , there exists a sequence of sets  $\{Q_i\}_i : \mathbb{P}(Q_i) \rightarrow 1$  such that*

$$-\eta + \gamma'_o \tilde{Z}_T < \mathbb{E}(\hat{Y}_T^{REG} - Y_T)^2 \mathbf{1}_{Q_i} < \eta + \gamma'_o \tilde{Z}_T.$$

If the number of target zones contained in one source zone  $S_{n,i}$  is bounded, the error at source level can be approximated by  $\mathbb{E}(Y_{n,i})$  and hence because  $Var(Y_T) = \mathbb{E}(Y_T) = \gamma'_o \tilde{Z}_T$ , this theorem says that the quadratic prediction error of the regression predictor is asymptotically equivalent to the variance of the underlying process. In the same conditions, the relative error at source level can be approximated by

$$(3.30) \quad \text{Re}_{n,i}^{REG} \approx \frac{1}{\sqrt{\mathbb{E}(Y_{n,i})}} = \frac{1}{\sqrt{\alpha_o k_n |S_{n,i}| + \beta_o k_n x_{n,i}}}.$$

Equation (3.30) shows that the relative error of the regression predictor is going to be small when the number of points on a source zone is large. However, this number being bounded by condition (C1), this relative error cannot converge to zero in this framework.

Let us now turn attention to the difference between the relative prediction errors of the Poisson regression method and that of the areal weighting and the dasymetric methods. If the target zones are nested within the source zones and the number of target zones contained in one source is bounded, we get the following approximation at source level for the differences between the relative errors of the methods when  $\mathbb{E}(Y_{n,i})$  are large and when  $\sum_T (\frac{|T|}{|S_{n,i}|} - \frac{|x_t|}{|x_{n,i}|})^2$  is not too small:

$$(3.31) \quad [(\text{Re}_{n,i}^{REG})^2 - (\text{Re}_{n,i}^{DAW})^2] \approx -\frac{1}{4}(1 - \Delta_{n,i})^2 \sum_T (\frac{|T|}{|S_{n,i}|} - \frac{x_T}{x_{n,i}})^2$$

$$(3.32) \quad [(\text{Re}_{n,i}^{REG})^2 - (\text{Re}_{n,i}^{DAX})^2] \approx -\frac{1}{4}(1 + \Delta_{n,i})^2 \sum_T (\frac{|T|}{|S_{n,i}|} - \frac{x_T}{x_{n,i}})^2.$$

This result shows that, among the three methods: areal weighting, dasymetric and Poisson regression, regression outperforms the other two methods asymptotically (negative sign). However, from the proof in the annex, we can see that if  $(\frac{|T|}{|S_{n,i}|} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}}) = 0$  then the regression is less accurate than areal weighting and dasymetric asymptotically so that none of them is always dominant.

The difference between the accuracy of the regression method and the other two methods depends on the difference of ratios  $\frac{|T|}{x_T} - \frac{|S|}{x_S}$ : the higher this difference, the larger the difference between regression and the other two.

The fact that the regression predictor doesn't satisfy the pycnophylactic property is not a surprise but the fact that it does satisfy this property on the whole region is interesting. The idea of scaling to obtain the pycnophylactic property can be found also in Yuan et al. (1997) for ordinary linear regression without theoretical justifications; we have extended it to the Poisson regression case and provided some theoretical motivation for it.

We now turn attention to the scaled regression and prove it is better than the unscaled one and that its accuracy can be approximated by that of the unfeasible oracle predictor.

The first lemma proves an asymptotic equivalence between the scaled regression predictor and the unfeasible oracle predictor.

**Lemma 3.3.5.** *For any target  $T$ ,*

$$(3.33) \quad \hat{Y}_T^{ScR} - \hat{Y}_T^C \rightarrow_p 0.$$

The next result is about the quadratic prediction error of the scaled Poisson regression predictor.

**Theorem 3.3.6.** For any  $\eta > 0$ , there exists a sequence of sets  $\{Q_i\}_i : \mathbb{P}(Q_i) \rightarrow 1$  such that

$$-\eta + \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o} < \mathbb{E}(\hat{Y}_T^{ScR} - Y_T)^2 \mathbf{1}_{Q_i} < \eta + \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o}.$$

Since  $\text{Er}_T^C = \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o}$ , this theorem shows that the quadratic prediction error of the scaled regression predictor is asymptotically equivalent to the one of the oracle. Consequently, the scaled regression method is the best among the areal weighting, the dasymetric and the regression predictors. In the same conditions as in equation (3.30), one can derive the following approximation for the relative error at source level

$$(3.34) \quad \text{Re}_S^{ScR} \approx \frac{1}{\sqrt{\mathbb{E}(Y_S)}} \sqrt{1 - \frac{\sum_{t \in s} \mathbb{E}(Y_T)^2}{\mathbb{E}(Y_S)^2}},$$

and we see that  $\sqrt{1 - \frac{\sum_{t \in s} \mathbb{E}(Y_T)^2}{\mathbb{E}(Y_S)^2}}$  measures the relative asymptotic efficiency between the regression and the scaled regression. This result says that the larger gain of scaling will be obtained in situations where the number of targets is small and they have heterogeneous sizes.

### 3.4 Simulated toy example

We devise a simple simulation to illustrate these results. On a square region  $\Omega$  with  $16 \times 16$  cells, we build three systems of sources with respectively 4, 14, and 64 sources (see Figure 3.1). We simulate two Poisson point processes (our auxiliary information) with an expected overall number of points of 100,000:  $X_1$  is very inhomogeneous (Gini coefficient of cell counts of 0.74 with 100,247 points) and  $X_2$  is very homogeneous (Gini coefficient of cell counts of 0.03 with 100,008 points).

Target variables are then generated following model (3.2). For each of the auxiliary variables, we choose three couples of coefficients  $\alpha, \beta$  to study the effects of imbalance so that we get six different target variables. In order to study the asymptotic effect, we multiply the basic auxiliary information by 4 in the case of 14 sources and by 16 in the case of 64 sources so that the ratio  $k_n/n = |S_{n,i}| / (n |S_{n,1}|)$  is approximately independent of the number of sources  $n$ .

We then apply the four considered methods (areal weighting, dasymetric, Poisson regression and scaled Poisson regression) to transfer the data from each of the three systems of source zones to cell level which our target level here. For each case, we generate the data 1000 times, and calculate relative prediction errors for each method and each iteration. Table 3.1 reports the median relative errors in the case of 14 sources for auxiliary information  $X_1$  and  $X_2$  for some choices of coefficients  $\alpha$  and  $\beta$ .

Figure 3.2 shows the impact of the homogeneity, expected number of points and imbalance on the difference between the error of DAW and the error of DAX. On the left panel, targets from source  $S_1$  with the same level of imbalance and total expected number of points are shown and we see that, in accordance with theorem (3.2.4), the difference between the two errors increases with the level of inhomogeneity (solid line), and in accordance with equation (3.22), as  $D$  increases, the ratio gets larger (dashed line), then stabilizes to its limiting value. The central panel shows similarly the impact of  $\mathbb{E}(Y_S)$  with the circles corresponding to targets of source  $S_3$  (circles, imbalance of  $-0.8$  and 23801 points) and source  $S_1$  (triangles, imbalance of  $-0.87$  and 8073 points): the triangles are below the circles. In the right panel where circles belong to source  $S_3$  and triangles to source  $S_2$  (imbalance of  $-0.72$  and 23938 points), we may select targets with the



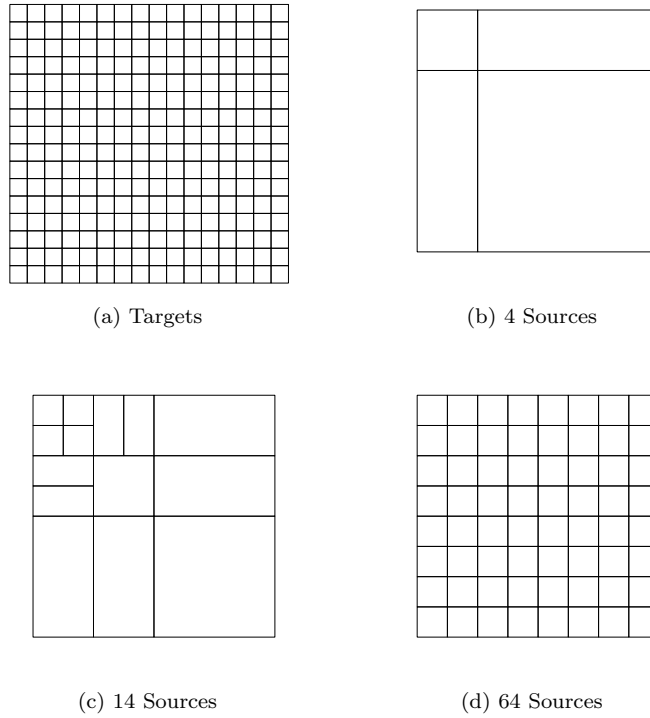


Figure 3.1: Spatial polygons for targets and sources

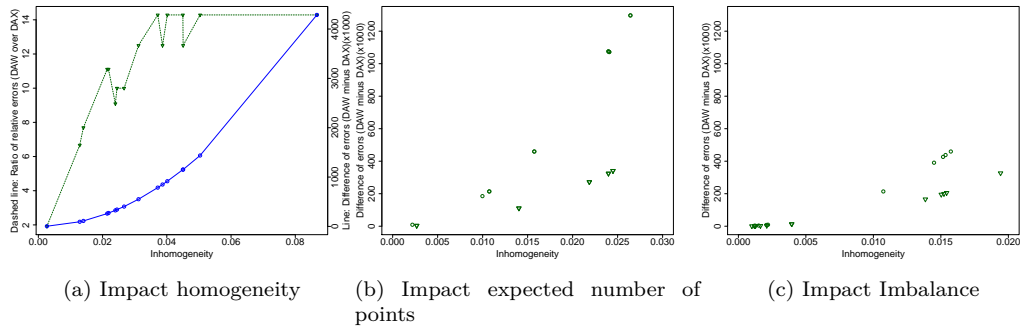


Figure 3.2: Comparison of DAW and DAX

same level of inhomogeneity and see that triangles are below circles due to the absolute value of the imbalance.

Table 3.2 reports the median relative errors in the case of 4, 14 and 64 sources for auxiliary information  $X_2$  and two sets of coefficients. The two tables also present the mean of the target variables at target, source and also region level (because it appears in Theorem 3.3.4) and the

<b>MET</b>	<b>Level</b>	<b><math>\mathbb{E}(Y)</math></b>	<b>DAW</b>	<b>DAX</b>	<b>REG</b>	<b>ScR</b>	<b>Oracle</b>	<b>% Pos.</b>	<b><math>\Delta_S</math></b>
$\alpha = 100$ $\beta = 1$ $X_1$	Target	634	11.950	25.383	3.966	3.911	3.925	56	-0.924
	Source	25654	14.548	2.614	0.617	0.553	0.555	25	-0.738
	Region	503388	5.224	1.268	0.140	0.132	0.132		0.963
$\alpha = 600$ $\beta = 1$ $X_1$	Target	2634	4.233	47.301	1.936	1.906	1.920	56	-0.616
	Source	48278	10.230	10.048	0.455	0.422	0.420	25	-0.050
	Region	1015388	2.591	3.752	0.099	0.095	0.095		0.994
$\alpha = 1000$ $\beta = 0.1$ $X_1$	Target	4023	1.589	69.052	1.559	1.537	1.549	100	0.597
	Source	34907	1.597	16.289	0.531	0.497	0.501	100	0.876
	Region	1064099	0.265	5.966	0.097	0.094	0.094		1
$\alpha = 100$ $\beta = 1$ $X_2$	Target	1960	3.435	2.317	2.261	2.223	2.213	0	-0.627
	Source	16059	1.339	0.798	0.787	0.740	0.738	0	-0.596
	Region	502432	0.278	0.150	0.142	0.138	0.137		-0.585

Table 3.1: **Relative prediction errors for the case with 14 sources (in percentages).**

theoretical oracle prediction error as a benchmark (see Theorem 3.3.6). The last column of the two tables report the imbalance at source level  $\Delta_S$ . At last the second to last column of Table 3.1 reports the percentage of targets (first line) and the percentage of sources (second line) with a positive imbalance.

Let us first use Table 3.1 to study the influence of the imbalance  $\Delta_S$ . Concerning the sign of  $\Delta_S$ , in the case  $\alpha = 1000, \beta = 0.1$  and auxiliary information  $X_1$ , the sign of  $\Delta_S$  according to theory should be favorable to the DAW method, which turns out to be true at all levels (target, source, region). Similarly, in the case  $\alpha = 100, \beta = 1$  and auxiliary information  $X_2$ , the sign is favorable to the DAX method and indeed it is true at all levels. If we now look at the case  $\alpha = 100, \beta = 1$  and auxiliary information  $X_1$ , the situation is more contrasted since the median imbalance is negative but the maximum is positive. Indeed there are four sources on which the imbalance is positive but they contain 56% of the targets and the result is that DAW is better at target level whereas DAX is better at source and region levels. If we now compare the first two sets of parameters, for  $\alpha = 100, \beta = 1$ , the median imbalance is negative with a large absolute value whereas for  $\alpha = 600, \beta = 1$ , the median imbalance is negative but small in absolute value. Consequently the theory is in agreement with the data in the first case but in disagreement in the second one: one should not derive conclusions when  $|\Delta_S|$  is small.

The influence of the overall number of points  $\mathbb{E}(Y_\Omega)$ , linked to the number of sources  $n_S$  by  $k_n$ , is clear from both tables: the relative errors are decreasing functions of this number.

With Table 3.2, we can see the influence of the homogeneity of the auxiliary information. We recall that  $X_2$  is rather homogeneous and  $X_1$  inhomogeneous and indeed the relative errors for DAW and DAX are smaller for  $X_2$  than for  $X_1$  due to the term  $D$  from equation (3.21). Note that this influence is much less when one compares the two regression methods. Let us now turn attention to comparing the dasymetric methods and the regression ones. In the case  $\alpha = 100, \beta = 1$  and auxiliary information  $X_2$ , we see in Table 3.2 that for 64 sources the basic regression is less good than DAX at target level and we can explain it by the fact that  $\alpha$  is small. However most of the time, the regression methods outperform DAX and DAW. In the case  $\alpha = 1000, \beta = 0.1$ , DAW is best for 4 and 14 sources but gets worse than the scaled

MET	$n$	$E(Y)$	DAW	DAX	REG	ScR	Oracle	$\Delta_S$
$\alpha = 100$ $\beta = 1$ $X_2$	4	490	5.313	4.572	4.547	4.534	4.4815	-0.604
		23874	0.850	0.656	0.654	0.650	0.641	-0.598
		125608	0.377	0.287	0.286	0.284	0.280	-0.587
	14	1960	3.435	2.317	2.261	2.223	2.213	-0.627
		16059	1.339	0.798	0.787	0.740	0.738	-0.596
		502432	0.278	0.150	0.142	0.138	0.137	-0.585
	64	7840	2.806	1.188	1.129	0.983	0.976	-0.627
		31280	1.648	0.655	0.565	0.493	0.490	-0.591
		2009728	0.230	0.083	0.071	0.062	0.061	-0.568
$\alpha = 1000$ $\beta = 0.1$ $X_2$	4	1039	3.087	4.476	3.133	3.122	3.087	0.922
		49907	0.444	0.804	0.453	0.450	0.443	0.924
		266000	0.193	0.361	0.196	0.196	0.192	0.926
	14	4156	1.528	3.577	1.559	1.530	1.527	0.916
		33285	0.516	1.398	0.548	0.515	0.513	0.924
		1064035	0.095	0.307	0.097	0.095	0.094	0.926
	64	16624	0.686	3.275	0.776	0.673	0.672	0.916
		66488	0.344	1.955	0.387	0.337	0.336	0.925
		4256013	0.043	0.272	0.048	0.042	0.042	0.93

Table 3.2: **Relative prediction errors (in percentages) for auxiliary information  $X_2$ .**

regression for 64 sources. If we now compare the basic regression to the scaled regression, we can compute from Table 3.2 the relative efficiency of the scaled regression with respect to the basic regression and we see that it increases with the number of sources and reaches a value of around 13% for 64 sources. Note that the scaled regression is very comparable to the benchmark. Concerning the difference between the relative errors of dasymetric and the regression given by equation (3.31), we see in Table 3.1 that for  $X_1$ , the term  $\sum_T \left( \frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2$  is large (because  $X_1$  is inhomogeneous) and that indeed the regression is always much better than DAX and DAW). In the case of  $X_2$ , the term  $\sum_T \left( \frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2$  is small and one must then look at the influence of the other term  $(1 + \Delta_S)^2$  or  $(1 - \Delta_S)^2$ . In the case  $\alpha = 100, \beta = 1$ ,  $(1 + \Delta_S)^2$  is quite large and thus the regression is better than DAX. In the case  $\alpha = 1000, \beta = 0.1$ ,  $(1 - \Delta_S)^2$  is small and the regression is worse than DAW.

We now turn attention to the robustness of the methods with respect to the model. As previously with the same geometrical design, we generate two auxiliary information scenarios:  $X_1$  is as in the previous simulation, and  $X_3$  is inhomogeneous and uncorrelated with  $X_1$  (correlation coefficient of  $-0.16$ ). A target variable  $Y$  is generated from  $X_3$  with the relationship  $Y_A \sim \mathcal{P}(600|A| + X_3)$ . We transfer  $Y$  from the first set of 14 sources to the cells (Figure 3.1) by using areal weighting interpolation, dasymetric interpolation with  $X_1$  and  $X_3$  as auxiliary variables, the regression methods (REG and SCR) with the true model (areal effect and  $X_3$ ), a simple model with only the areal effect, an auxiliary variable model with an irrelevant variable (with area and  $X_1$ ), an auxiliary variable model involving an unnecessary variable (the area and both  $X_1$  and  $X_3$ ). Table 3.3 presents the results.

The most accurate method is the scaled regression with area and  $X_3$  (true model). Note that the relative error for DAW and ScR with area only is the same which was expected since we proved that in that case the two methods coincide. The regression methods for the model

Methods	Relative error
DAW	7.74
DAX with $X_3$	9.49
REG with area and $X_3$	2.66
ScR with area and $X_3$	2.62
DAX with $X_1$	14.70
REG with area and $X_1$	10.48
ScR with area and $X_1$	8.26
REG with area	10.62
ScR with area	7.74
REG with area, $X_1$ and $X_3$	2.66
ScR with area, $X_1$ and $X_3$	2.62

Table 3.3: **Robusness of methods.**

involving area plus  $X_1$  and  $X_3$  as auxiliary have the same errors (2.66% and 2.62%): in other words using unnecessary variables in the regression does not decrease the accuracy. On the other hand, if we use the regression with a wrong choice of auxiliary variable, it gives bad predictions (10.48% and 8.26% for the model with area and  $X_1$ , 10.62% and 7.74% for the model with only areal effect). The dasymetric method with  $X_3$  is better than with  $X_1$  (9.49% vs 14.70%) which makes sense because the correlation of the target variable  $Y$  with  $X_3$  is 0.998 while with  $X_1$  it is of  $-0.159$  however we see that despite the strong correlation between  $Y$  and  $X_3$  the dasymetric method with  $X_3$  is not so good because the areal effect is strong. The scaled regression is always better than the regression method and the scaled regression in the case of areal effect model yields the same result as the areal weighting interpolation method. Finally, the difference between scaled and unscaled regression is larger when the wrong auxiliary variable is used as one can see comparing the regressions using area and  $X_1$  and the regressions using area and  $X_3$ .



## Chapter 4

# Application and R-package

In this chapter, we present an application with a very rich data set UScensus2000. Since the database provides many information about counts at several levels of zonal systems, it fits with the conditions we have studied in Chapter 3. The application is not only illustrating the theoretical results in Chapter 3 but also leading to many new questions. Besides comparisons between the dasymetric and regression based methods according to the selection of auxiliary variables, the size and relative size of the spatial supports (the source and the target zones), we also consider comparing two types of regressions: Poisson and gaussian by using the transformation proposed in the Chapter 2. Our application is done using functions from an R-package presented in Section 4.2. The package provides all simple methods corresponding to those in Table 2.1.

### 4.1 Application

#### 4.1.1 Data

The UScensus2000 database contains data from the US decennial census at several different geographic levels (in particular: states, counties, tracts, block groups and blocks). The package contains functions for aggregating the demographic information at any of these levels. It is therefore highly adapted to test areal interpolation techniques. Since all considered variables are available at all geographical levels, we will be able to assess the accuracy of the considered interpolation methods based on the true target values on the selected target zones.

Following Almquist et al. (2010), we choose to work with the target variable corresponding to the number of house owners in a given zone for the extensive case. We also select a corresponding intensive variable which is the percentage of house owners, the weights being given by the number of households in the given zone. As potential auxiliary information, we use the covariates presented in Almquist (2010) which are the number (resp: percentage) of non hispanic white, of non hispanic black, of non hispanic asian, of hispanic, of married households with children in the population. The first four percentages are with respect to the population whereas the last one is with respect to the number of households. As far as spatial scale is concerned, we decide to use three different scenarios in the state of Ohio. The first scenario is the disaggregation of the target variable from county level (source) to tract level (target) for the whole of the state of Ohio. Ohio has 88 counties and 2941 tracts so that on average one county contains 33 tracts. The second and third scenario use the county of Franklin as the whole region. Franklin counts 284 tracts, 887 block groups and 22826 blocks. The second scenario is the disaggregation of the

target variable from tracts (source) to block groups (targets) in the county of Franklin. In this case, one tract contains on average 3 block groups. The third scenario is the disaggregation of the target variable from tracts (source) to blocks (targets) and in that case one tract contains on average 80 blocks. A particular feature of these scenarios is that in all cases the target zones are nested within the source zones.

We first perform some exploratory analysis of the variables at source and at target levels. At the county level (source) for the whole Ohio, all extensive variables are strongly positively correlated whereas the corresponding intensive variables are much less correlated and display some negative correlations. For example the percentage of white is negatively strongly correlated with the percentage of black (-0.97) but the percentage of hispanic has no clear linear relationship with other intensive variables. At the tract level (source) on Franklin county, the correlations are smaller. More precisely, the number of house owners is still strongly positively correlated with population, number of households, number of whites and married households with children, but not clearly correlated with number of blacks, number of asians and hispanic. For corresponding intensive variables, correlations are smaller than 0.4 except for percentage of white and percentage of blacks which are strongly negatively correlated (-0.97). On Franklin, correlations at tract level are very similar to correlations at block group level (target).

Besides the scaled regression method for the Poisson case mentioned in the previous chapter, we propose here to do the same for gaussian regression of intensive variables by using the empirical conditional expectation of  $Y_t$  given the source values, i.e. for  $t \subset S$

$$(4.1) \quad \hat{Y}_t^{SCL} = \mathbb{E}(Y_t \mid \widehat{Y_1, \dots}, Y_S) = \hat{Y}_t^{REG} - \hat{Y}_s^{REG} + Y_s,$$

where  $\hat{Y}_t^{REG}$  and  $\hat{Y}_s^{REG}$  are the fitted values for  $Y_t$  and  $Y_s$  respectively and where  $\hat{Y}_s^{REG}$  is simply given by the aggregation rule  $\hat{Y}_s^{REG} = \sum_{t \subset S} w_{t:t \subset S} \hat{Y}_t^{REG}$ .

Another approach for this problem is to use an EM-algorithm strategy as done in Flowerdew and Green (1993) for the Poisson case and in Flowerdew and Green (1993) for the Gaussian case. Indeed the areal interpolation can be considered as a missing data problem with target values as missing data. We can summarize the steps as follows: the expectation step (E-step) is either (3.28) or (4.1) and yield values for the targets and the maximization step is the regression at target level based on models mentioned in Section 2.3.3.

## 4.1.2 Results

Table 4.1 summarizes the notations used for presenting the results. To illustrate the meaning of this table, let us take two examples. The indices have two or three positions: for example “Dhh” or “I.Dhh”. When necessary, an additional index in position one will indicate either the intensive/extensive nature or the spatial support depending upon background.

In positions 2 and 3 (potentially after the dot), an index “Dhh” means that we are using the *dasymeric* method (D) with the auxiliary information *percentage of households* (hh). An index “I.Dhh” specifies moreover that it is for the *intensive* target variable (percentage of house owner). In positions 2 and 3, an index “Ebe” means that we are using the regression method with the *EM* algorithm (E) for the *best* model choice (be) (independent variables are chosen by AIC criteria). An index “B.Ebe” specifies moreover that it is for the with *blocks* as target zones.

At source level, the criterion for evaluating the quality of methods is the relative error of prediction

$$e_s = \frac{\sqrt{\sum_{t \subset S} (\hat{Y}_t - Y_t)^2}}{Y_s}.$$

Meaning		Notation	Index Position
Dependent Variables	Intensive	I.	1
	Extensive	E.	
Spatial support	Block	B.	1
	Block group	Bg.	
Methods	Dasymetric	D	2
	Regression	R	
	Scaled regression	S	
	EM	E	
Independent variables	number/percentage of white	w	3
	number/percentage of black	b	
	number/percentage of asian	a	
	number/percentage of hispanic	h	
	number/percentage of married with children	m	
	population	p	
	households	hh	
	area	aa	
	full (all variables)	f	
	best variable choice	be	

Table 4.1: Notations

### Auxiliary information selection

The choice of a good auxiliary information is an important question for areal interpolation in practice. First of all, it is unclear whether a choice of variables which is good for the regression step will be the best for predicting target values. On the other hand, it is difficult to devise a prediction-targeted criterion adapted to this situation: indeed, one does not observe any target value hence it is not straightforward to extend cross-validation to this case. By lack of a better alternative, we have chosen to use a variable choice strategy based on the AIC criterion. Note this selection has been performed using the R package *MASS* (Ripley et al. (2015)) for gaussian regression and the R package *glmulti* (Calcagno et al. (2010)) for Poisson regression.

We compare several dasymetric methods obtained by using the different auxiliary information at our disposal using scenario 1 (Ohio) for the extensive case. Table 4.2 displays the corresponding median error criterions showing that it is very important to select the best auxiliary information since the relative error can vary from around 3 percent to 30 percent. The second row displays the correlations and the non monotonicity of these numbers shows that one should not trust correlation to select an auxiliary information. Figure 4.1 presents the boxplots of the source errors for the best dasymetric (here: based on the number of whites), the worst dasymetric (here: the number of blacks) and an intermediate case corresponding to areal weighting interpolation. We see that not only the best choice outperforms the other ones by far but also that the variability across sources is quite high for these choices.

### Comparison between different regression methods

In this section, we focus on comparing the different methods using scenario 1 (Ohio) for the extensive case. The implementation of the Poisson regression approach presents some peculiarities. The first one is about the choice of link function. The usual choice for Poisson regression is



	Dw	Dm	Dp	Dhh	Dh	Daa	Da	Db
error	2.76	3.10	3.14	3.75	14.97	21.95	22.87	31.31
correlation	0.99	0.99	1	1	0.88	0.02	0.93	0.95

Table 4.2: Performance of the dasymetric method for each auxiliary information

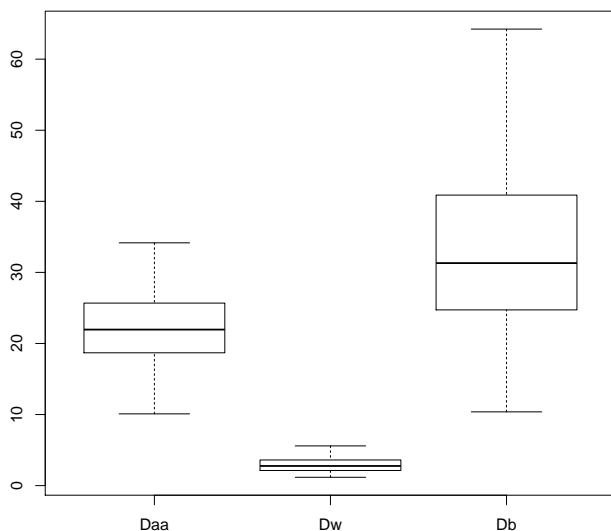


Figure 4.1: Dasymetric methods for Ohio - extensive approach

the logarithm link leading to  $\mathbb{E}(Y) = \exp(\sum_{i=1}^p \beta_i X_i)$  and it is the so called natural link in this generalized linear model. However we argue in Do et al. (2014b) that the identity link is more adapted when relating such extensive variables to auxiliary extensive variables. For example, it seems more natural that the *number of house owners* is proportional to the *population* rather than to be exponentially related to the population. Moreover, empirically, the AIC criterion is 1000 times bigger for the log link specification.

The second one is about the constant term. With the identity link, it does not make sense to include a constant in such a model because a constant is not an extensive variable.

Figure 4.2 presents the boxplots of the counties error criterions for the state of Ohio and for the Poisson regression performed on the *number of house owners*. Table 4.3 presents the corresponding median error criterions.

The order of magnitude of the errors is around 3 percent and they are very comparable. The selection of variables strategy selects the model without the variables *married household with children and area* but it seems that keeping a full model does not make a big difference. We

Sbe	Ebe	Sf	Ef	Rbe	Rf
3.069	3.069	3.193	3.196	3.482	3.594

Table 4.3: Median error criterions - Poisson regressions for Ohio

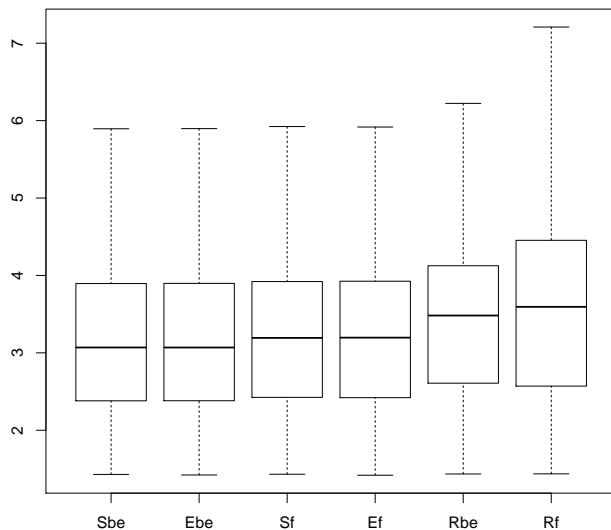


Figure 4.2: Poisson regression methods for Ohio - extensive approach

see that the scaled regression tends to perform better in general, and even better than the EM approach. However, it turns out that the best of the regression methods gets a 3.069 error criterion and does not outperform the best dasymetric obtained in the previous section with a 2.764 error criterion.

### Intensive versus extensive approach

For the purpose of interpolating the target variable *number of house owners*, we have the choice between two strategies. The first one is to work on the raw variable which is extensive and use a Poisson regression approach. The second one is to work on the *percentage of house owners*, use a gaussian regression approach and transform back the predicted percentages into counts using the knowledge of population on targets if known. In a different situation when this knowledge is not guaranteed, a more complex method is available which disaggregates separately numerator and denominator of this percentage using extensive variables methods. In this section we compare the first two approaches only, the last one giving results very similar to the second one in our case. We use scenarios 1 and 2. For Ohio in scenario 1, Figure (4.3) shows that the best method is the dasymetric method with auxiliary information given by the *number of whites* applied to the count target variable *number of house owners*. For Franklin in scenario 2, the right panel of Figure 4.6 shows again that the best result is obtained when working with the count variable rather than the percentage and it is obtained by the regression on the best subset of auxiliary variables. We also see that scaled gaussian regression that we introduced in section 4.1.1 is the second best.

### 4.1.3 Spatial scale

In this section, we examine the effect of spatial scale on the areal interpolation problem. For this we compare scenarios 2 and 3 on the county of Franklin. Figure 4.4 (respectively Figure 4.5) presents the distributions across sources of the error criterion in the case of disaggregation of the

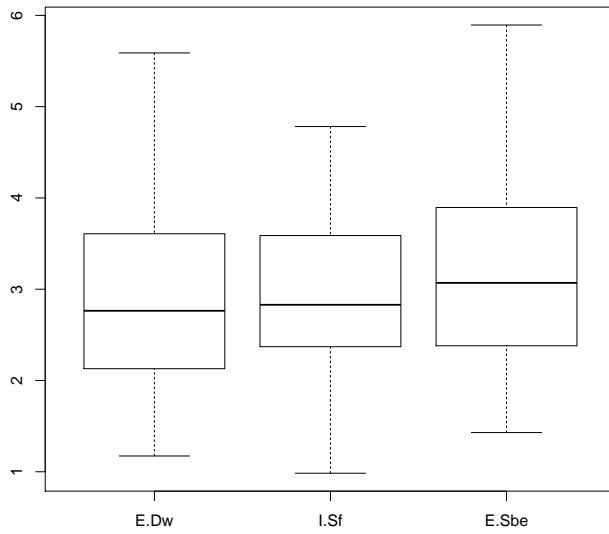


Figure 4.3: Best methods for Ohio - intensive and extensive approaches

extensive variable *number of house owners* (respectively of the intensive variable *percentage of house owners*) at block level and at block group level. Tables 4.4 and 4.5 display the corresponding median error criteria.

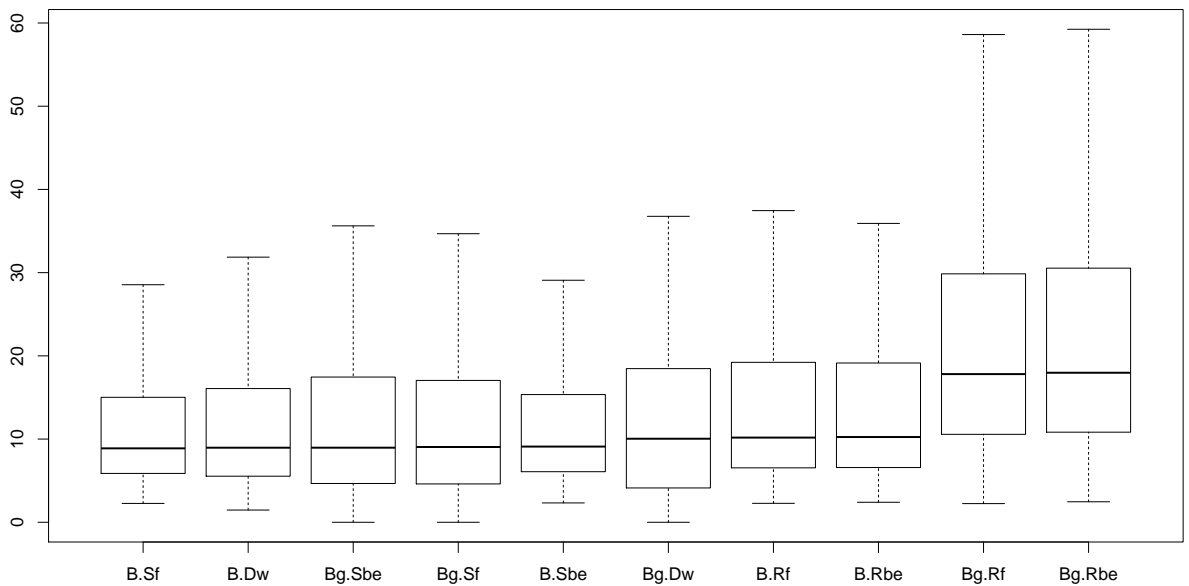


Figure 4.4: Methods for the extensive case for Franklin

For the extensive case, the best model strategy selects the model without the variables *number*

B.Sf	B.Dw	Bg.Sb	Bg.Sf	B.Sb	Bg.Dw	B.Rf	B.Rb	Bg.Rf	Bg.Sb
8.884	8.962	8.964	9.041	9.097	10.041	10.178	10.246	17.808	17.969

Table 4.4: Performance of the methods in the extensive case

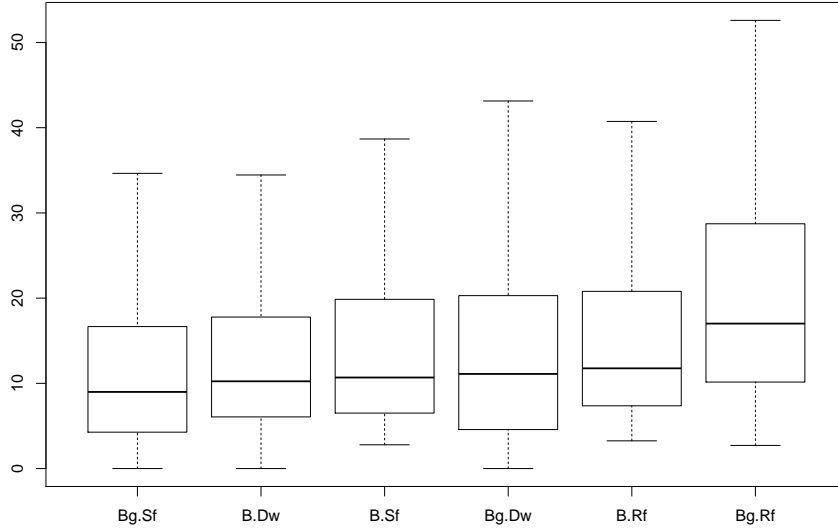


Figure 4.5: Methods for the intensive case for Franklin

of blacks and area. For the intensive case, the best model is the full model.

We note that the best accuracy for Franklin is around 10% whereas it is around 3% for Ohio. The two situations are difficult to compare because even though the number of targets per source is 33 for Ohio and is between 3 for scenario 2 on Franklin and 80 for scenario 3 on Franklin, on the other hand, there are larger numbers of house owners on the sources of Ohio than the sources of Franklin and the sizes of sources and targets are different.

When we compare scenarios 2 and 3 on Figure 4.6, we see that disaggregation to blocks is more accurate than to blockgroups. Even though the second problem seems easier because the blockgroups are coarser than blocks, one should not forget that the auxiliary information is used at target level resulting in a larger amount of information used for blocks. The medians of source error criterions at block level are thus slightly smaller and the variances are much smaller.

Finally, it turns out that the scaled regression methods always outperform the unscaled ones and that the improvement is stronger at block group level than at block level because the information is poorer at block group level. Indeed before scaling the regression methods at block group level were much worse than at block level and the scaling almost wipes off this difference.

Figures 4.7 and 4.8 present map of Franklin county. Figure 4.7 shows the errors of the three

Bg.Sf	B.Dw	B.Sf	Bg.Dw	B.Rf	Bg.Rf
8.993	10.24	10.684	11.102	11.761	17.013

Table 4.5: Performance of the methods in the intensive case

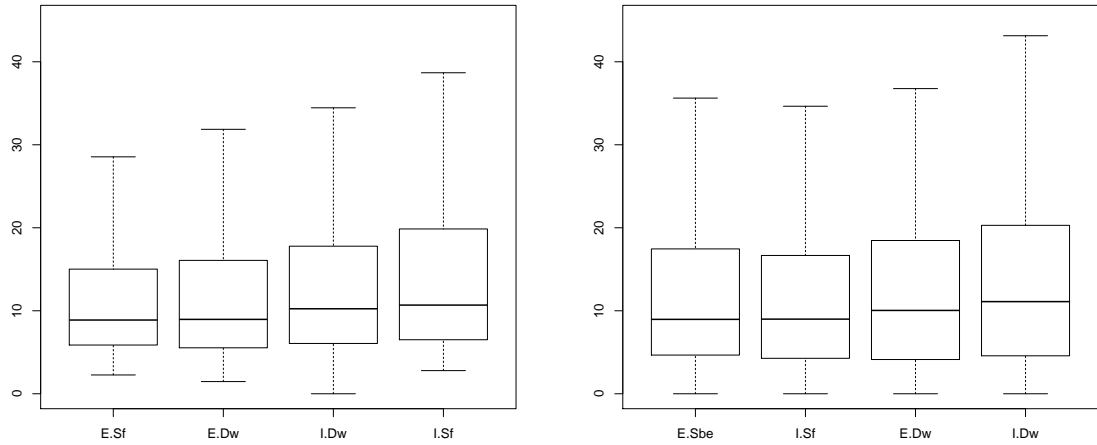


Figure 4.6: Comparison of block (left) and blockgroup (right) levels for Franklin - intensive and extensive variables

methods at source level (284 tracts). The two methods: DAX with the auxiliary variable "numbers of white people" (left panel) and the scaled regression (right panel) seem equivalent whereas the worst methods among the three ones is the regression (center panel). We also see that larger errors locate around the center of Franklin where tracts are smaller. This fact might indicate spatial dependence which we need to study more in future work.

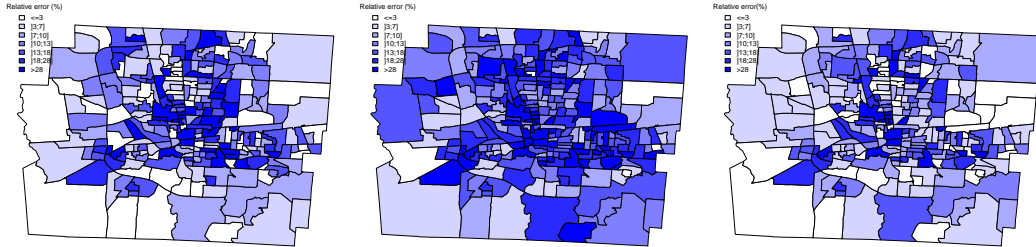


Figure 4.7: Errors on Franklin county: DAX with white (left), Regression (center) and scaled regression (right)

A focus on the center of Franklin is displayed in Figure 4.8 that presents the true number of house owners at target level (upper left panel), the disaggregation with the DAX method using the number of white inhabitants as auxiliary information (upper right panel, the choice of this particular variable has been optimized), the disaggregation with a regression (obtained after selecting the best set of auxiliary information variables including number of whites, blacks, hispanics and married couples, in the lower left panel) and the scaled regression with the same auxiliary information on the lower right panel. The three disaggregations seem satisfactory except for one tract for which the prediction gets negative for the two regression methods and is not good for the DAX either despite being positive.

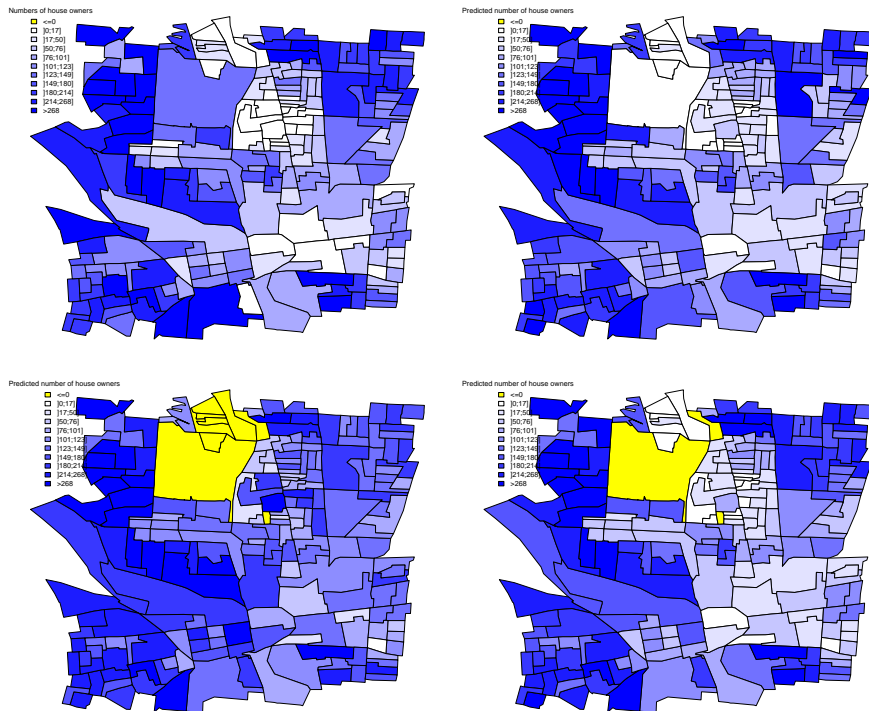


Figure 4.8: Center of Franklin county: True counts (upper left), DAX with white (upper right), Regression (lower left) and scaled regression (lower right)

## 4.2 Package

Even though the need for interpolating data is big, there is no easy package to implement these available methods. This package aims to give simple functions which correspond to the classified methods and enable practitioners without deep knowledge about statistics or mathematics to select and utilize a suitable method based on their available data.

The package includes the functions of the methods presented in the table 2.1. They are named *daw*, *dax*, *dax2step*, *rwo*, *reg* corresponding to the areal weighting, original dasymetric, 2-step dasymetric (dasymetric with control zones), regression without auxiliary information and regression with auxiliary information. In addition, the scaled regression is also programmed with the name *scr*.

Each function depends on several inputs: First point, we need to provide the spatial supports: sources, targets, and controls (if available), then define the target variable  $Y$ , the auxiliary variables  $X$  (if necessary), the nature of these variables (extensive or intensive), the weights for each intensive variables (if any - areal as default). The format for the sources must be a `SpatialPolygonsDataFrame` with observations of  $Y$ . The format for the targets and controls must be either `SpatialPolygonsDataFrame` or `SpatialPolygons`. The target variable is obviously defined on the sources. Spatial support of the auxiliary variables differs according to the methods: *dax* requires the availability of  $X$  at the intersection level  $A_{st}$  which is quite strict. In general, the *dax2step* is more often used instead because it loosens condition about the spatial support

of  $X$ . In order to implement *dax2step*, the controls containing the values of  $X$  are necessarily provided. On the other hand, the spatial level of  $X$  for the case *reg* is quite flexible: it might be the target, the control or the intersection zones. However, *scl* needs availability of  $X$  on all the intersection zones. We hence use the same technique in the *dax2step* to ease the constraint: apply *daw* to interpolate  $X$  into the intersections if necessary. The choice of algorithms for each method is determined by the nature of the variables  $Y, X$  so the nature is necessary to be set. We recall all intensive variables are linked to their weights, the default weights are the ones defined by the area zone. When all the variables are set, the functions are simply programmed based on the formulae in the previous chapters.

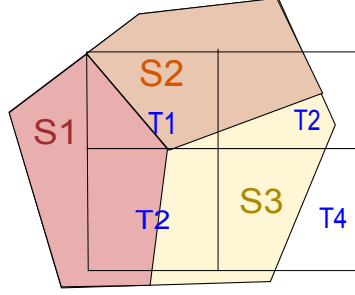


Figure 4.9: *scaling* option case

There is an option named "scaling" which points out if "zone-scale" step is used. Zone-scale is a technique which is offered when the union of the target zones  $\mathbf{T} = \cup_t T_t$  and the union of the sources  $\mathbf{S} = \cup_s S_s$  don't coincide (see Figure 4.9). Indeed, when those unions are different, the aggregation formulae

$$(4.2) \quad Y_s = \sum_{t:t \cap s \neq \emptyset} Y_{st}$$

and

$$(4.3) \quad Y_t = \sum_{s:s \cap t \neq \emptyset} Y_{st}$$

are no longer correct if

$$S_s \neq \cup_{t:t \cap s \neq \emptyset} A_{st}$$

and

$$T_t \neq \cup_{s:s \cap t \neq \emptyset} A_{st}.$$

Because the aggregation on the sources (4.2) is never used for the areal weighting and dasy-metric method, so the *scaling* option of *daw*, *dax* functions is applied for the targets which are not nested within the union  $\mathbf{S}$ .

Precisely, for the areal weighting interpolation method, if one believes that the un-overlapped zone (the part of  $T_t$  outside  $\mathbf{S}$ ) has similar property as the overlapped one ( $T_t \cap \mathbf{S}$ ), we propose a zone-scale step as follows: for the extensive case

$$(4.4) \quad \hat{Y}_t^{DAW} = \frac{|T_t|}{|T_t \cap \mathbf{S}|} \hat{Y}_{T_t \cap \mathbf{S}}^{DAW}$$

and for the intensive case, we keep the same value on  $T_t \cap \mathbf{S}$ .

If one believes that the target variable is null on the region outside  $\mathbf{S}$  (for instance if the target variable is population or population density and the region outside  $\mathbf{S}$  is in the ocean, hence that region is not populated), the extensive case predictors are kept, but the intensive one is necessarily scaled as follow: if the weights  $Z$  of the intensive target variable  $Y$  are available

$$(4.5) \quad \hat{Y}_t^{DAW} = \frac{Z_{t \cap \mathbf{S}}}{Z_t} \hat{Y}_{t \cap \mathbf{S}}^{DAW}$$

where  $Z_{t \cap \mathbf{S}} = \sum_{s: s \cap t \neq \emptyset} Z_{st}$ . When  $Z$  is not available, 4.5 is replaced by

$$(4.6) \quad \hat{Y}_t^{DAW} = \frac{|T_t \cap \mathbf{S}|}{|T_t|} \hat{Y}_{T_t \cap \mathbf{S}}^{DAW}$$

We emphasize that the *scaling* option is *yes* when the neighbor of  $\mathbf{S}$  is believed to be similar to  $\mathbf{S}$  and *no* if there is a "zero" value on that adjacent region. The default is *yes*.

The case of the regression methods is more complicated because the aggregation steps (4.2), (4.3) are both necessary. We hence use zone-scale technique in order to simplify the regression equation by the following process: when  $\mathbf{S} \neq \mathbf{T}$ , instead of implementing the regression on the original systems of zones, we restrict the implementation on a new system of zones which are the projections of the originals on  $\mathbf{U} = \mathbf{S} \cap \mathbf{T}$ . More precisely

$$(4.7) \quad T'_t = T_t \cap \mathbf{U} \text{ and } S'_s = S_s \cap \mathbf{U}$$

We then apply zone-scale if the neighbor of  $\mathbf{U}$  is similar to itself in order to get value of  $Y$  on the new source zones and to recover the original target zones after the regression.

## Functions

---

*daw* Areal weighting interpolation method

---

### Description

Function that predicts the values of the target variable on target zones by the areal interpolation method.

### Usage

```
daw<-function(sources, targets, y, nature="extensive", scaling=TRUE)
```

### Argument

sources	a SpatialPolygonsDataFrame containing the source polygons as well as the values of the target variable on the source zones
targets	a SpatialPolygons or a SpatialPolygonsDataFrame
y	target variable name
nature	nature of target variable (extensive and intensive)
scaling	zone-scale option



## Value

Vector of target variable values on target zones.

---

*dax* Dasymetric method

---

## Description

Function that predicts the values of the target variable on the target zones by the dasymetric method when an auxiliary variable is available on the intersection zones.

## Usage

```
dax<-function(sources, targets, y, st.df, x, Z.y=NULL, Z.x=NULL, scaling=TRUE)
```

## Argument

sources	a SpatialPolygonsDataFrame containing the source polygons as well as the values of the target variable on the source zones
targets	a SpatialPolygons or a SpatialPolygonsDataFrame
y	target variable name
st.df	a data.frame containing the values of the auxiliary variable on the intersection zones
x	auxiliary variable name
Z.y	weights for an intensive target variable. NULL corresponds to the extensive case
Z.x	weights for an intensive auxiliary variable. NULL corresponds to the extensive case
scaling	zone-scale option

## Value

Vector of target variable values on target zones.

---

*dax.2step* 2-step dasymetric method

---

## Description

Function that predicts the values of the target variable on the target zones by the dasymetric method when an auxiliary variable is available on the control zones.

## Usage

```
dax.2step<-function(sources, targets, controls, y, x, Z.y=NULL, Z.x=NULL, algo="1daw",  
scaling="none")
```

### Argument

<code>sources</code>	a <code>SpatialPolygonsDataFrame</code> containing the source polygons as well as the values of the target variable on the source zones
<code>targets</code>	a <code>SpatialPolygons</code> or a <code>SpatialPolygonsDataFrame</code>
<code>controls</code>	a <code>SpatialPolygons</code> or a <code>SpatialPolygonsDataFrame</code> containing auxiliary variable
<code>y</code>	target variable name
<code>x</code>	auxiliary variable name
<code>Z.y</code>	weights for an intensive target variable. NULL corresponds to the extensive case
<code>Z.x</code>	weights for an intensive auxiliary variable. NULL corresponds to the extensive case
<code>algo</code>	number of <i>daw</i> step used
<code>scaling</code>	zone-scale option

### Value

Vector of target variable values on target zones.

---

*rwo* Regression without auxiliary information method.

---

### Description

Function that predicts the values of the target variable on target zones by the regression based method when there is no auxiliary variable.

### Usage

```
rwo<-function(sources, targets, y, nature="extensive", scaling=TRUE)
```

### Argument

<code>sources</code>	a <code>SpatialPolygonsDataFrame</code> containing the source polygons as well as the values of the target variable on the source zones
<code>targets</code>	a <code>SpatialPolygons</code> or a <code>SpatialPolygonsDataFrame</code>
<code>y</code>	target variable name
<code>nature</code>	nature of target variable (extensive and intensive)
<code>scaling</code>	zone-scale option

### Value

Vector of target variable values on target zones.



## Chapter 5

# Perspective and 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 2.1.

We have not addressed in the 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.

In Chapter 3 we have analyzed the accuracy of four areal interpolation methods: areal weighting interpolation, dasymetric interpolation, Poisson regression and scaled Poisson regression for the case of count data. We have introduced a model based on an underlying Poisson point pattern to be able to evaluate the accuracy of the different methods. We have proposed a scaled version of the Poisson regression method resulting in the enforcement of the pycnophylactic property. Areal weighting interpolation and dasymetric interpolation have been compared with a finite distance approach and the regression methods have been compared together and with the previous ones with an asymptotic approach.

We found out that one shouldn't rely on the correlation of the target variable and the auxiliary variable or on the homogeneity of the target variable to decide between areal interpolation or dasymetric but we should also take into account the relative imbalance between the areal effect and the auxiliary effect. A strong areal effect leads to the dominance of the areal weighting interpolation and a strong auxiliary effect is in favor of the dasymetric method. Moreover, the imbalance index allows to approximate the ratio of the two relative errors and their lower bounds as the number of points on the source zones gets large. We establish the formula for the best linear predictor (therefore better than the areal weighting and the dasymetric), which leads to the introduction of the scaled regression method.

For the comparison of areal weighting and dasymetric, a combination of several factors explains

the complexity of the behavior: the size of sources, the auxiliary information, the number and size of target zones, . . . The error at source level is better when sources are divided into a smaller number of target zones. A large number of points makes the error at source level worse but improves the accuracy of the relative error. These two types of errors have the same behavior as a function of the imbalance index. The impact of the expected number of points and of the inhomogeneity on the comparative advantage of the methods should not be forgotten: indeed when we have several sources, the sign of the imbalance index may vary from source to source and the overall effect, being an aggregate of the source level effect, will also depend on the magnitude of the source error differences which is driven by the expected number of points and by the inhomogeneity. We proved that the accuracy of the unfeasible composite predictor is decreasing when the expected number of points are similar on all targets and this fact extends to scaled regression (due to the approximation results).

To be able to include the regression methods in the comparison, we need to resort to some asymptotic approach. We propose an asymptotic framework and prove that the Poisson regression prediction error is equivalent to the variance of the underlying process and for the scaled regression, it is approximated by the composite's prediction error. These results show the regression predictor is not automatically better than the areal weighting interpolation or the dasymetric method, but when the number of points at source level is large, it is in general better. Finally the scaled regression turns out to be the best one among the considered methods. These results are confirmed by our simulation study of the last section. The robustness with respect to the model is also considered. The simulations show that a model with extra auxiliary variables doesn't create any loss while missing variables or unrelated variables (in place of the correct ones) decrease the accuracy of all methods.

We used the very rich database from R-package UScensus2000 to illustrate the theoretical methods in Chapter 4. We should keep in mind that this study has a particular geometry due to the nesting of targets into sources. In a more general case, some border effects will interplay but we believe that, as long as the size of targets is much smaller than the size of sources (disaggregation), the results should not be very different.

We would like to emphasize the three main conclusions of this study. About the choice of auxiliary variable for the dasymetric method we have seen that the performance can vary wildly from one choice to another so this choice is crucial. The second one is that sometimes dasymetric can be better than scaled regression which means that it might be more important to select one good auxiliary information rather than throwing a lot of weakly related variables in the regression. The last one is that scaled regression is very close to the EM algorithm (and much simpler) and often even better.

Because of the lack of an easy implemented package about areal interpolation methods, we started programming an R-package including all the most basic and popular methods. There is a detailed guideline that helps users to determine easily properties of their data that lead to the choice of appropriate methods. The functions also cover the practical case which is displayed by the option *scaling* where some targets contain zones outside the union of the sources.

There remain many open questions. We already considered the extensive variable case using a Poisson point pattern model and got many interesting results. A similar approach for the intensive case can be studied. Instead of using a Poisson point process, we think it can be possible to use a marked point process to analyze the accuracy of intensive variables. In addition, we have been working with parametric models. We think one potential approach might be the semi-parametric models. Indeed, our target attribute might have parametric relationship with some variables and non-parametric relationship with other variables. In a future research, the assumption about the existence of an underlying process is kept but instead of modeling the process's intensity to be linearly correlated, we assume it depends on two types of auxiliary

variables: one group is linearly linked and the other has non-parametric connection with the target variable. There is another question for the areal interpolation problem which is the selection of variables. Usual criteria are difficult to adapt here because prediction is done for a statistical unit (target) different from the unit used in the regression step. Works needs to be done to look for an appropriate tool to choose a good variable for the dasymetric methods or a good set of variables for the regression methods.



## Chapter 6

# Appendix

We present in the appendix all proofs for the dasymetric method, those of the areal weighting interpolation method use the same arguments.

### Proof of Lemma 3.2.1 and Lemma 3.2.2

From (3.10) and the properties of a Poisson point process we have

$$\mathbb{E}(\hat{Y}_T^{DAX} - Y_T) = \mathbb{E}\left(\frac{x_T}{x_S} Y_S - Y_T\right) = \frac{x_T}{x_S}(\alpha|S| + \beta x_S) - (\alpha|T| + \beta x_T) = \alpha|S|\left(\frac{x_T}{x_S} - \frac{|T|}{|S|}\right).$$

Taking into account the independence of two disjoint target zones with the fact that the target  $T$  is a portion of the source  $S$  the variances of each method are given as follows

$$\begin{aligned} \text{Var}(\hat{Y}_T^{DAX} - Y_T) &= \text{Var}\left(\frac{x_T}{x_S} Y_S - Y_T\right) = \frac{x_T^2}{x_S^2} \text{Var}(Y_S) + \text{Var}(Y_T) - 2\frac{x_T}{x_S} \text{Cov}(Y_S, Y_T) \\ &= \frac{x_T^2}{x_S^2} \mathbb{E}(Y_S) + \mathbb{E}(Y_T) - 2\frac{x_T}{x_S} \text{Var}(Y_T) = \alpha|S|\left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 + \beta x_T \left(1 - \frac{x_T}{x_S}\right) + \alpha|T|\left(1 - \frac{|T|}{|S|}\right). \end{aligned}$$

Summing up the variances at target level with the fact that  $\sum_T \frac{|T|}{|S|} = \sum_T \frac{x_T}{x_S} = 1$ , we get the variances at source level in Lemma 3.2.2

### Proof of Theorem 3.2.3

From Lemma 3.2.1 and the fact that  $\alpha|S| = I_S(|\cdot|)\mathbb{E}(Y_S)$ ,  $\beta X_S = I_S(X)\mathbb{E}(Y_S)$  we have

$$\begin{aligned} \text{Er}_T^{DAW} &= I_S(X)\mathbb{E}(Y_S)\left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 + I_S(X)\mathbb{E}(Y_S)\left(\frac{x_T}{x_S} - \frac{x_T^2}{x_S^2}\right) + I_S(|\cdot|)\mathbb{E}(Y_S)\left(\frac{|T|}{|S|} - \frac{|T|^2}{|S|^2}\right) \\ &\quad + I_S(X)^2\mathbb{E}(Y_S)^2\left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 \\ \text{Er}_T^{DAX} &= I_S(|\cdot|)\mathbb{E}(Y_S)\left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 + I_S(X)\mathbb{E}(Y_S)\left(\frac{x_T}{x_S} - \frac{x_T^2}{x_S^2}\right) + I_S(|\cdot|)\mathbb{E}(Y_S)\left(\frac{|T|}{|S|} - \frac{|T|^2}{|S|^2}\right) \\ &\quad + I_S(|\cdot|)^2\mathbb{E}(Y_S)^2\left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2. \end{aligned}$$



If the expectation of the number of points is sufficiently large, we can approximate the ratio of the two errors (the relative errors) as follows

$$\frac{\text{Er}_T^{DAW}}{\text{Er}_T^{DAX}} \approx \frac{I_S(X)^2}{I_S(|\cdot|)^2}, \quad \frac{\text{Re}_T^{DAW}}{\text{Re}_T^{DAX}} \approx \frac{I_S(X)}{I_S(|\cdot|)}.$$

At source level, we get a similar result by adding up errors on all target zones using the fact that  $\sum_T \frac{|T|}{|S|} = \sum_T \frac{x_T}{x_S} = 1$  :

$$\begin{aligned} \text{Er}_S^{DAX} &= I_S(|\cdot|)\mathbb{E}(Y_S) \sum_T \left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 + I_S(X)\mathbb{E}(Y_S)\left(1 - \sum_T \frac{x_T^2}{x_S^2}\right) + I_S(|\cdot|)\mathbb{E}(Y_S)\left(1 - \sum_T \frac{|T|^2}{|S|^2}\right) \\ &\quad + I_S(|\cdot|)^2\mathbb{E}(Y_S)^2 \sum_T \left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 \\ \Rightarrow \text{Re}_S^{DAX} &= \frac{1}{\mathbb{E}(Y_S)} \left[ I_S(|\cdot|)\left(1 - \sum_T \frac{x_T}{x_S}\right)^2 + I_S(X)\left(1 - \sum_T \frac{x_T^2}{x_S^2}\right) + I_S(|\cdot|)\left(1 - \sum_T \frac{|T|^2}{|S|^2}\right) \right] \\ &\quad + I_S(|\cdot|)^2 \sum_T \left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2. \end{aligned}$$

Using the relationship  $I_S(|\cdot|) + I_S(X) = 1$ , the above results prove Theorem 3.2.3.

#### Proof of Theorem 3.2.4

Lemma 3.2.1 yields

$$\begin{aligned} \text{Er}_T^{DAW} - \text{Er}_T^{DAX} &= \text{Var}(\hat{Y}_T^{DAW} - Y_T) + [\mathbb{E}(\hat{Y}_T^{DAW} - Y_T)]^2 - \text{Var}(\hat{Y}_T^{DAX} - Y_T) - [\mathbb{E}(\hat{Y}_T^{DAX} - Y_T)]^2 \\ &= \left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 \frac{(\beta x_S - \alpha|S|)}{(\beta x_S + \alpha|S|)} \left( (\beta x_S + \alpha|S| + 1)(\beta x_S + \alpha|S|) \right) = \left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2 \Delta_S (\mathbb{E}(Y_S) + 1) \mathbb{E}(Y_S). \end{aligned}$$

#### Proof of Theorem 3.2.5

We calculate the error of the oracle predictors then minimize with respect to  $w$  to find the optimal  $w^*$

$$\begin{aligned} \hat{Y}_T^C &= w\hat{Y}_T^{DAW} + (1-w)\hat{Y}_T^{DAX} = \left[w\frac{|T|}{|S|} + (1-w)\frac{x_T}{x_S}\right]Y_S := uY_S \\ \text{Bias}_T^2 &= [\mathbb{E}(\hat{Y}_T^{DAW} - \hat{Y}_T^{DAX})]^2 = (u\mu_S - \mu_T)^2 \\ \text{Var}_T &= \text{Var}(uY_S - Y_T) = u^2\mu_S + \mu_T - 2u\mu_T \\ \text{Er}_T &= u^2\mu_S(\mu_S + 1) - 2u\mu_T(\mu_S + 1) + \mu_T^2 + \mu_T \\ u^* &= \underset{u}{\text{argmin}} \text{Er}_T = \frac{\mu_T}{\mu_S} \Leftrightarrow w^* = \frac{\alpha|T|}{\alpha|S| + \beta x_S}. \end{aligned}$$

Substituting the  $w^*$  in (3.23) we get the oracle predictor (3.24).

The bias, variance and error of the above oracle predictor are calculated as follows

$$\begin{aligned} \text{Bias} &= \mathbb{E}(\hat{Y}_T^C - Y_T) = 0 \\ \text{Er}_T^C &= \text{Var}(\hat{Y}_T^C - Y_T) = \text{Var}\left(\frac{\mu_T}{\mu_S} Y_S - Y_T\right) = \frac{x_T^2}{x_S^2} \mu_S + \mu_T - 2 \frac{x_T}{x_S} \mu_T - \mu_S \left(\frac{x_T}{x_S} - \frac{\mu_T}{\mu_S}\right)^2 \\ &= \text{Var}(\hat{Y}_T^{DAX} - Y_T) - \mu_S \left(\frac{x_T}{x_S} - \frac{\mu_T}{\mu_S}\right)^2 = \text{Var}(\hat{Y}_T^{DAW} - Y_T) - \mu_S \left(\frac{|T|}{|S|} - \frac{\mu_T}{\mu_S}\right)^2. \end{aligned}$$

Since  $Y_T|Y_S \sim \text{Bi}(Y_S, \frac{\mathbb{E}(Y_T)}{\mathbb{E}(Y_S)})$ , we have  $\mathbb{E}(Y_T|Y_S) = \frac{\mathbb{E}(Y_T)}{\mathbb{E}(Y_S)} Y_S = \hat{Y}_T^C$ . This shows that the oracle predictor is the best linear predictor.

### Proof of Theorem 3.3.1

It is easy to see that  $\mathbb{E}_\gamma(s_n(\gamma)) = 0$ ,  $\mathbb{E}_\gamma(H_n(\gamma)) = F_n(\gamma)$ . We further simplify the notations and use  $s_n, F_n, H_n, \mathbb{E}$  instead of  $s_n(\gamma_o), F_n(\gamma_o), H_n(\gamma_o), \mathbb{E}_{\gamma_o}$ . It is clear that the matrix  $H_n$  is positive definite and therefore the log likelihood function is concave which leads to a unique minimum. In the sequel, we also need the square root  $F_n^{1/2}$  of the symmetric matrix  $F_n$ , i.e.  $F_n^{1/2} F_n^{1/2} = F_n$ . To prove the theorem, we will prove the following lemmas

**Lemma 6.0.1.** *Under conditions (C1) and (C2), the normed score function  $F_n^{-1/2} s_n$  is asymptotically normal*

$$(6.1) \quad F_n^{-1/2} s_n \rightarrow_d \mathcal{N}(0, \mathbf{I}).$$

**Lemma 6.0.2.** *Under conditions (C1) and (C2), for all  $\delta > 0$*

$$(6.2) \quad \max_{\gamma \in N_n(\delta)} \|V_n(\gamma) - \mathbf{I}\| \rightarrow_p 0,$$

where  $N_n(\delta) = \{\gamma : \|F_n^{1/2}(\gamma - \gamma_o)\| \leq \delta\}$ ,  $V_n(\gamma) = F_n^{-1/2} H_n(\gamma) F_n^{-1/2}$ .

Lemma 6.0.1 is proved by using the Lindeberg-Feller theorem.

Indeed, for  $\tau$  fixed with  $\tau' \tau = 1$ , considering the triangular array

$$z_{n,i} = \tau' F_n^{-1/2} \frac{\tilde{Z}_{n,i}}{\gamma' \tilde{Z}_{n,i}} (y_{n,i} - \gamma' \tilde{Z}_{n,i})$$

we have  $\mathbb{E}(z_{n,i}) = 0$ ,  $\sum_i \text{Var}(z_{n,i}) = 1$ . We will show that the Lindeberg condition is satisfied, i.e. for any  $\varepsilon > 0$

$$(6.3) \quad \sum_i \mathbb{E}(z_{n,i}^2 \mathbf{1}_{|z_{n,i}| > \varepsilon}) \rightarrow 0$$

as  $n \rightarrow \infty$ .

Let  $a_{n,i} = \tau' F_n^{-1/2} \frac{\tilde{Z}_{n,i}}{\gamma' \tilde{Z}_{n,i}}$ , because  $z_{n,i}^2 = a_{n,i}^2 (y_{n,i} - \gamma' \tilde{Z}_{n,i})^2$ ,  $\mathbb{E}(z_{n,i}^2 \mathbf{1}_{|z_{n,i}| > \varepsilon}) = a_{n,i}^2 \mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^2 \mathbf{1}_{|y_{n,i} - \gamma' \tilde{Z}_{n,i}| > \frac{\varepsilon}{|a_{n,i}|}})$ , yields

$$\begin{aligned} \sum_i \mathbb{E}(z_{n,i}^2 \mathbf{1}_{|z_{n,i}| > \varepsilon}) &= \sum_i a_{n,i}^2 \mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^2 \mathbf{1}_{|y_{n,i} - \gamma' \tilde{Z}_{n,i}| > \frac{\varepsilon}{|a_{n,i}|}}) \\ &\leq \left(\sum_i a_{n,i}^2\right) \sup_i \mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^2 \mathbf{1}_{|y_{n,i} - \gamma' \tilde{Z}_{n,i}| > \frac{\varepsilon}{|a_{n,i}|}}). \end{aligned}$$

Moreover, condition (C1) yields that there is a positive number  $K_1$  s.t.  $\frac{1}{\gamma' \tilde{Z}_{n,i}} < K_1, \forall (n, i)$ , hence

$$\sum_i a_{n,i}^2 = \tau' F_n^{-1/2} \sum_i \frac{\tilde{Z}_{n,i} \tilde{Z}'_{n,i}}{(\gamma' \tilde{Z}_{n,i})^2} F_n^{-1/2} \tau < K_1 \tau' F_n^{-1/2} \sum_i \frac{\tilde{Z}_{n,i} \tilde{Z}'_{n,i}}{\gamma' \tilde{Z}_{n,i}} F_n^{-1/2} \tau = K_1.$$

In addition, conditions (C1) (C2) lead to  $\max_i \frac{\varepsilon}{|a_{n,i}|} \rightarrow \infty$  as  $n \rightarrow \infty$ , hence for any  $M > 0, \exists n_1$  s.t.  $\forall n > n_1$

$$\begin{aligned} & \sup_i \mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^2 \mathbf{1}_{|y_{n,i} - \gamma' \tilde{Z}_{n,i}| > \frac{\varepsilon}{|a_{n,i}|}}) \leq \sup_i \mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^2 \mathbf{1}_{|y_{n,i} - \gamma' \tilde{Z}_{n,i}| > M}) \\ & \leq \sup_i \sqrt{\mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^4 \mathbf{1}_{|y_{n,i} - \gamma' \tilde{Z}_{n,i}| > M})} \leq \sup_i \sqrt{\mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^4 \frac{\text{Var}(y_{n,i} - \gamma' \tilde{Z}_{n,i})}{M^2})} \\ & = \sup_i \sqrt{\gamma' \tilde{Z}_{n,i} (1 + 3\gamma' \tilde{Z}_{n,i}) \frac{\gamma' \tilde{Z}_{n,i}}{M^2}} < \frac{K_2}{M}. \end{aligned}$$

Hence  $\sup_i \mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^2 \mathbf{1}_{|y_{n,i} - \gamma' \tilde{Z}_{n,i}| > \frac{\varepsilon}{|a_{n,i}|}}) \rightarrow 0$  as  $n \rightarrow \infty$ , where the existence of  $K_2$  is derived from condition (C1). The argument yields the (6.3). So Lemma 6.0.1 holds.

*Proof of Lemma 6.0.2*

Using the same notation in the proof of Lemma 6.0.1,  $\tau$  fixed s.t.  $\tau' \tau = 1$ , let  $b_{n,i} = \tau' F_n^{-1/2} \tilde{Z}_{n,i}$ , the equation (6.2) can be rewritten as

$$(6.4) \quad \tau'(V_n(\gamma) - \mathbf{I})\tau = A_n + B_n + C_n,$$

where  $A_n = \sum_i b_{n,i}^2 (\frac{1}{(\gamma' \tilde{Z}_{n,i})^2} - \frac{1}{(\gamma'_o \tilde{Z}_{n,i})^2}) (y_{n,i} - \gamma' \tilde{Z}_{n,i})$ ,  $B_n = \sum_i b_{n,i}^2 \frac{1}{(\gamma'_o \tilde{Z}_{n,i})^2} (y_{n,i} - \gamma' \tilde{Z}_{n,i})$ ,  $C_n = \sum_i b_{n,i}^2 (\frac{1}{\gamma' \tilde{Z}_{n,i}} - \frac{1}{\gamma'_o \tilde{Z}_{n,i}})$ .

We will prove that the three terms converge in probability to 0 as  $n$  tends to  $\infty$ . To prove the convergence of  $B_n$ , we first study its properties. We have

$$\begin{aligned} \mathbb{E}(B_n) &= 0 \\ \text{Var}(B_n) &= \sum_i b_{n,i}^4 \frac{1}{(\gamma'_o \tilde{Z}_{n,i})^4} \text{Var}(y_{n,i} - \gamma' \tilde{Z}_{n,i}) = \sum_i b_{n,i}^4 \frac{1}{(\gamma'_o \tilde{Z}_{n,i})^4} \gamma' \tilde{Z}_{n,i} \\ &\leq \sum_i b_{n,i}^2 \frac{1}{\gamma'_o \tilde{Z}_{n,i}} \sup_i \frac{b_{n,i}^2}{(\gamma'_o \tilde{Z}_{n,i})^3} \gamma' \tilde{Z}_{n,i} = \sup_i b_{n,i}^2 \frac{1}{(\gamma'_o \tilde{Z}_{n,i})^3} \gamma' \tilde{Z}_{n,i} < K_3 \sup_i b_{n,i}^2. \end{aligned}$$

Because of the boundedness of  $(\gamma'_o \tilde{Z}_{n,i})^3$  and the definition of  $N_n(\delta)$ ,  $\gamma' \tilde{Z}_{n,i}$  is bounded when  $n$  is large enough, moreover,  $\sup_i b_{n,i}^2 \rightarrow 0$  due to the condition (C1) (C2), therefore  $B_n \rightarrow_p 0$ .

We can use similar argument to prove  $A_n \rightarrow_p 0, C_n \rightarrow 0$ , and this shows that the lemma 6.0.2 holds.

### Proof of Proposition 3.3.2

Let  $z_{nij} \sim \mathcal{P}(\gamma'_o Z_{n,i}) - \gamma'_o Z_{n,i} := \tilde{z}_{n,i}, j = 1, 2, \dots, k_n$  i.i.d .

This yields  $\sum_j z_{nij} = Y_{n,i} - \gamma'_o \tilde{Z}_{n,i}$ . We have  $\mathbb{E}(z_{nij}) = 0, \sum_j \text{Var}(z_{nij}) = \gamma'_o \tilde{Z}_{n,i}$ .

We will prove that this array satisfies the Lindeberg-Feller condition, i.e.  $\forall \delta > 0, \sum_j \mathbb{E}(z_{nij}^2 \mathbf{1}_{|z_{nij}| > \delta}) \rightarrow 0$ , as  $n \rightarrow \infty$ . Indeed,  $\sum_j \mathbb{E}(z_{nij}^2 \mathbf{1}_{|z_{nij}| > \delta}) = k_n \mathbb{E}(\tilde{z}_{n,i}^2 \mathbf{1}_{|\tilde{z}_{n,i}| > \delta}) = \mathbb{E}(u_{n,i}^2 \mathbf{1}_{|u_{n,i}| > \sqrt{k_n} \delta})$ , where

$u_{n,i} = \sqrt{k_n} \tilde{z}_{n,i}$ . Because  $\mathbb{E}u_{n,i} = 0$ ,  $\mathbb{E}u_{n,i}^2 = \text{Var}u_{n,i} = k_n \text{Var}\tilde{z}_{n,i} = \gamma'_o \tilde{Z}_{n,i} < \infty$ . Moreover  $k_n \rightarrow \infty$  as  $n \rightarrow \infty$ , we have  $\mathbb{E}(u_{n,i}^2 \mathbf{1}_{|u_{n,i}| > \sqrt{k_n} \delta}) \rightarrow 0$  as  $n \rightarrow \infty$ .

From the Lindeberg-Feller theorem we get  $\frac{Y_{n,i} - \gamma'_o \tilde{Z}_{n,i}}{\sqrt{\gamma'_o \tilde{Z}_{n,i}}} \rightarrow_d \mathcal{N}(0, 1)$ . This proof can be applied at the target level, i.e.  $\frac{Y_T - \gamma'_o \tilde{Z}_T}{\sqrt{\gamma'_o \tilde{Z}_T}} \rightarrow_d \mathcal{N}(0, 1)$ .

### Proof of Theorem 3.3.3

The pycnophylactic property of the scaled regression predictor is obvious. To prove the pycnophylactic property of the regression predictor at region level, we sum up regression predictors over source zones

$$\hat{Y}_\Omega^{REG} = \sum_i \sum_{T: T \subset S_{n,i}} \hat{Y}_T^{REG} = \sum_i \sum_{T: T \subset S_{n,i}} \hat{\gamma}_n \tilde{Z}_T = \hat{\gamma}'_n \tilde{Z}_\Omega.$$

Recall that  $\hat{\gamma}$  is the solution of the score equation  $s_n(\gamma) = 0$ , i.e.

$$\sum_{i=1}^n \frac{\tilde{Z}_{n,i}}{\hat{\gamma}' \tilde{Z}_{n,i}} y_{n,i} - \tilde{Z}_{n,i} = 0 \Rightarrow \sum_{i=1}^n \frac{\hat{\gamma}' \tilde{Z}_{n,i}}{\hat{\gamma}' \tilde{Z}_{n,i}} y_{n,i} - \hat{\gamma}' \tilde{Z}_{n,i} = 0 \Leftrightarrow \sum_{i=1}^n y_{n,i} - \hat{\gamma}' \tilde{Z}_\Omega = 0 \Leftrightarrow \hat{\gamma}' \tilde{Z}_\Omega = y_\Omega.$$

That is, the regression predictor satisfies the pycnophylactic property on the region  $\Omega$ .

To study the pycnophylactic property of the regression predictor at source level, we consider  $\hat{Y}_{n,i}^{REG} - Y_{n,i}$ . We have

$$\begin{aligned} \hat{Y}_{n,i}^{REG} - Y_{n,i} &= \hat{\gamma}'_n \tilde{Z}_{n,i} - Y_{n,i} = (\hat{\gamma}'_n - \gamma'_o) \tilde{Z}_{n,i} - (Y_{n,i} - \gamma'_o \tilde{Z}_{n,i}) \\ &= F_n^{-1/2} F_n^{1/2} (\hat{\gamma}'_n - \gamma'_o) \tilde{Z}_{n,i} - (Y_{n,i} - \gamma'_o \tilde{Z}_{n,i}). \end{aligned}$$

The first term converges to 0 in distribution due to the conditions (C1), (C2) and the theorem 3.3.1. The second term is different from 0, even asymptotically (Proposition 3.3.2).

Moreover, because of the boundedness of  $\tilde{Z}_{n,i}$ , the above argument yields the result of Theorem 3.3.3.

If  $\tilde{Z}_T$  is bounded below, a similar result at target level holds.

### Proof of Theorem 3.3.4

For any target  $T$ , the error of the regression predictor on the target is

$$\mathbb{E}(\hat{Y}_T^{REG} - Y_T)^2 = \mathbb{E}(\hat{\gamma}'_n \tilde{Z}_T - \gamma'_o \tilde{Z}_T)^2 + \mathbb{E}(\gamma'_o \tilde{Z}_T - Y_T)^2 - 2\mathbb{E}(\hat{\gamma}'_n \tilde{Z}_T - \gamma'_o \tilde{Z}_T)(\gamma'_o \tilde{Z}_T - Y_T).$$

From Theorem 3.3.1 and condition (C1), for any  $\eta_1 > 0$ ,  $\exists \varepsilon > 0$  s.t. when  $n$  is sufficiently large

$$(6.5) \quad \mathbb{E}(\hat{\gamma}'_n \tilde{Z}_T - \gamma'_o \tilde{Z}_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} < \eta_1$$

$$(6.6) \quad \|\mathbb{E}(\hat{\gamma}'_n \tilde{Z}_T - \gamma'_o \tilde{Z}_T)(\gamma'_o \tilde{Z}_T - Y_T) \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon}\| < \eta_1.$$

As we proved in Theorem 3.3.3  $(\hat{\gamma}'_n \tilde{Z}_T - \gamma'_o \tilde{Z}_T) \rightarrow_p 0$ , we have  $\mathbb{P}(\|\hat{\gamma}_n - \gamma_o\| < \varepsilon) \rightarrow 1$  as  $n \rightarrow \infty$ . In addition  $\mathbb{E}(\gamma'_o \tilde{Z}_T - Y_T)^2 = \gamma'_o \tilde{Z}_T$ . Hence there is  $n_1$  s.t.

$$\mathbb{E}(\gamma'_o \tilde{Z}_T - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| \geq \varepsilon} < \mathbb{E}(\gamma'_o \tilde{Z}_T - Y_T)^4 \mathbb{P}(\|\hat{\gamma}_n - \gamma_o\| \geq \varepsilon) < \eta_1$$

for  $n > n_1$ . In other words,  $\gamma'_o \tilde{Z}_T - \eta_1 < \mathbb{E}(\gamma'_o \tilde{Z}_T - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| > \varepsilon} < \gamma'_o \tilde{Z}_T$ .

This implies  $\forall \eta > 0$ ,  $\exists \varepsilon > 0$ ,  $n_1$  s.t. for  $n > n_1$

$$(6.7) \quad -\eta + \gamma'_o \tilde{Z}_T < \mathbb{E}(\hat{Y}_T^{REG} - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} < \eta + \gamma'_o \tilde{Z}_T$$

with a remark that  $\mathbb{P}(\|\hat{\gamma}_n - \gamma_o\| < \varepsilon) \rightarrow 1$  as  $n \rightarrow \infty$ .

Combining (6.5), (6.6), (6.7) we get Theorem 3.3.4.

### Proof of equations (3.31)

We rewrite the error of the areal interpolation and dasymetric for the asymptotic model. For a target  $T \subset S_{n,i}$ , from (3.25), (3.26), and Lemma 3.2.1 we have

$$\text{Er}_T^{DAX} = \gamma'_o \tilde{Z}_T - \frac{(\gamma'_o \tilde{Z}_T)^2}{\gamma'_o \tilde{Z}_{n,i}} + \gamma'_o \tilde{Z}_{n,i} \left( \frac{\tilde{x}_T}{\tilde{x}_{n,i}} - \frac{\gamma'_o \tilde{Z}_T}{\gamma'_o \tilde{Z}_{n,i}} \right)^2 + \alpha^2 \widetilde{|S_{n,i}|}^2 \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2.$$

A similar argument as in the proof of theorem 3.3.4 shows that, for any  $\eta_1 > 0, \varepsilon > 0, \exists n_1$  s.t.  $\forall n > n_1$

$$\text{Er}_T^{DAX} - \eta_1 < \mathbb{E}(\hat{Y}_T^{DAX} - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} < \text{Er}_T^{DAX}.$$

With  $\varepsilon$  chosen as in theorem 3.3.4, let  $Q_i = \{ \|\hat{\gamma}_n - \gamma_o\| < \varepsilon \}$ , we have

$$\gamma'_o \tilde{Z}_T - \text{Er}_T^{DAX} - \eta < \mathbb{E}(\hat{Y}_T^{REG} - Y_T)^2 \mathbf{1}_{Q_i} - \mathbb{E}(\hat{Y}_T^{DAX} - Y_T)^2 \mathbf{1}_{Q_i} < \gamma'_o \tilde{Z}_T - \text{Er}_T^{DAX} + \eta + \eta_1$$

for all  $n > n_1$ . Moreover,

$$\gamma'_o \tilde{Z}_T - \text{Er}_T^{DAX} = \beta \tilde{x}_{n,i} \frac{\tilde{x}_T^2}{\tilde{x}_{n,i}^2} + \alpha \widetilde{|S_{n,i}|} \frac{|\tilde{T}|^2}{\widetilde{|S_{n,i}|}^2} - \alpha \widetilde{|S_{n,i}|} \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2 - \alpha^2 \widetilde{|S_{n,i}|}^2 \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2.$$

Taking the sum over all target zones which belong to  $S_{n,i}$  then scaling the sum by  $\mathbb{E}(Y_{n,i})$  and calculating the differences in terms of  $\Delta_{n,i} = \Delta_{S_{n,i}}$ , we have

$$\begin{aligned} 4 \frac{\sum_T \gamma'_o \tilde{Z}_T - \text{Er}_T^{DAX}}{\mathbb{E}(Y_{n,i})^2} &= 4 \frac{1}{\mathbb{E}(Y_{n,i})} \left[ \frac{\beta \tilde{x}_{n,i}}{\mathbb{E}(Y_{n,i})} \sum_T \frac{\tilde{x}_T^2}{\tilde{x}_{n,i}^2} + \frac{\alpha \widetilde{|S_{n,i}|}}{\mathbb{E}(Y_{n,i})} \sum_T \frac{|\tilde{T}|^2}{\widetilde{|S_{n,i}|}^2} - \frac{\alpha \widetilde{|S_{n,i}|}}{\mathbb{E}(Y_{n,i})} \sum_T \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2 \right] \\ &\quad - 4 \left( \frac{\alpha \widetilde{|S_{n,i}|}}{\mathbb{E}(Y_{n,i})} \right)^2 \sum_T \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2 \\ &= 2 \frac{1}{\mathbb{E}(Y_{n,i})} \left[ (1 - \Delta_{n,i}) \sum_T \frac{\tilde{x}_T^2}{\tilde{x}_{n,i}^2} + (1 + \Delta_{n,i}) \sum_T \frac{|\tilde{T}|^2}{\widetilde{|S_{n,i}|}^2} - \right. \\ &\quad \left. - (1 - \Delta_{n,i}) \sum_T \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2 \right] - (1 + \Delta_{n,i})^2 \sum_T \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2. \end{aligned}$$

Using the same calculation for the areal weighting interpolation method, we have

$$\begin{aligned} 4 \frac{\sum_T \gamma'_o \tilde{Z}_T - \text{Er}_T^{DAW}}{\mathbb{E}(Y_{n,i})^2} &= 2 \frac{1}{\mathbb{E}(Y_{n,i})} \left[ (1 - \Delta_{n,i}) \sum_T \frac{\tilde{x}_T^2}{\tilde{x}_{n,i}^2} + (1 + \Delta_{n,i}) \sum_T \frac{|\tilde{T}|^2}{\widetilde{|S_{n,i}|}^2} - \right. \\ &\quad \left. - (1 - \Delta_{n,i}) \sum_T \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2 \right] - (1 + \Delta_{n,i})^2 \sum_T \left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right)^2. \end{aligned}$$

If  $\left( \frac{|\tilde{T}|}{\widetilde{|S_{n,i}|}} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right) = 0$  then the regression is less accurate than areal weighting and dasymetric asymptotically. If this difference increases, the difference between the regression and the other two methods gets smaller and then the regression method can do better than the other two methods.

Indeed, for example when  $\frac{x_T}{|T|} = \frac{\beta x_{S_{n,i}} - \alpha |S_{n,i}|}{2\beta |S|}$ , this yields  $T, x_T$  satisfy  $(\frac{|\widetilde{T}|}{|S_{n,i}|} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}}) \neq 0$ ,

we have  $\frac{(\gamma'_o \tilde{Z}_T)^2}{\gamma'_o \tilde{Z}_{n,i}} - \frac{1}{\gamma'_o \tilde{Z}_{n,i}} \beta^2 \tilde{x}_{n,i}^2 (\frac{|\widetilde{T}|}{|S_{n,i}|} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}})^2 = 0$ .

Therefore,  $\gamma'_o \tilde{Z}_T - \text{Er}_T^{DAW} = -\beta^2 \tilde{x}_{n,i}^2 (\frac{|\widetilde{T}|}{|S_{n,i}|} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}})^2 < 0$ .

Choosing  $\eta, \eta_1$  to be sufficient small, the regression predictor is asymptotically better than the areal weighting interpolation predictor. A similar result for the case of the dasymetric predictor can be proved similarly.

We therefore proved that none of the considered three methods is always dominant.

### Proof of Lemma 3.3.5

Assume  $T \in S_{n,i}$ , the difference between the predictors of scaled regression and oracle predictor is given by

$$\begin{aligned} \hat{Y}_T^{ScR} - \hat{Y}_T^C &= \frac{\hat{\gamma}'_n \tilde{Z}_T}{\hat{\gamma}'_n \tilde{Z}_{n,i}} Y_{n,i} - \frac{\gamma'_o \tilde{Z}_T}{\tilde{Z}_{n,i} \gamma_o} Y_{n,i} = (\hat{\gamma}'_n - \gamma'_o) \frac{\tilde{\gamma}'_n \tilde{Z}_{n,i} \tilde{Z}_T - \tilde{\gamma}'_n \tilde{Z}_T \tilde{Z}_{n,i}}{(\tilde{Z}_{n,i} \tilde{\gamma}'_n)^2} Y_{n,i} \\ &= (\hat{\gamma}'_n - \gamma'_o) F_n^{-T/2} F_n^{T/2} Y_{n,i} \frac{\tilde{\gamma}'_n \tilde{Z}_{n,i} \tilde{Z}_T - \tilde{\gamma}'_n \tilde{Z}_T \tilde{Z}_{n,i}}{(\tilde{Z}_{n,i} \tilde{\gamma}'_n)^2}, \end{aligned}$$

where  $\tilde{\gamma}'_n$  belongs to the segment of  $\hat{\gamma}'_n$  and  $\gamma'_o$ .

From Theorem 3.3.1, Proposition 3.3.2, and conditions (C1), (C2), we have

$$F_n^{T/2} (\hat{\gamma}'_n - \gamma'_o) \rightarrow_d \mathcal{N}(0, I), F_n^{-T/2} \frac{Y_{n,i} - \tilde{Z}_{S_{n,i}} \gamma_o}{\sqrt{\tilde{Z}_{S_{n,i}} \gamma_o}} \rightarrow_d 0, \frac{\tilde{\gamma}'_n \tilde{Z}_{n,i} \tilde{Z}_T - \tilde{\gamma}'_n \tilde{Z}_T \tilde{Z}_{n,i}}{(\tilde{Z}_{n,i} \tilde{\gamma}'_n)^2} \text{ bounded,}$$

In other words,  $\hat{Y}_T^{ScR} - \hat{Y}_T^C \rightarrow_p 0$ .

### Proof of Theorem 3.3.6

Because of the boundedness of  $\tilde{Z}_{n,i}$ , upper boundedness of  $\tilde{Z}_T$ , there exists

$$M = \sup_{\tilde{Z}_T, \tilde{Z}_{n,i}, \gamma \in B(\gamma_o, 1)} \left\| \frac{\tilde{Z}_{n,i} \gamma_n \tilde{Z}_T - \tilde{Z}_T \gamma_n \tilde{Z}_{n,i}}{(\tilde{Z}_{n,i} \gamma_n)^2} \right\| \mathbb{E}(Y_{n,i}^2),$$

where  $B(\gamma_o, 1) = \{\gamma : \|\gamma_o - \gamma\| < 1\}$ . Since  $\hat{\gamma}'_n - \gamma_o \rightarrow_p 0$ , the sequence  $\hat{\gamma}'_n, n = 1, 2, \dots$  is bounded, therefore for any  $\varepsilon > 0$ , when  $n$  is large enough

$$\sup_{\tilde{Z}_T, \tilde{Z}_{n,i}, \tilde{\gamma}'_n \in \text{segment}(\gamma_o, \hat{\gamma}'_n)} \left\| \frac{\tilde{Z}_{n,i} \tilde{\gamma}'_n \tilde{Z}_T - \tilde{Z}_T \tilde{\gamma}'_n \tilde{Z}_{n,i}}{(\tilde{Z}_{n,i} \tilde{\gamma}'_n)^2} \right\| \mathbb{E}(Y_{n,i}^2) \mathbf{1}_{\|\hat{\gamma}'_n - \gamma_o\| < \varepsilon} < M$$

For any  $\eta > 0$ , there is an  $\varepsilon > 0$  s.t.

$$\mathbb{E}(\hat{Y}_T^{ScR} - \hat{Y}_T^C)^2 \mathbf{1}_{\|\hat{\gamma}'_n - \gamma_o\| < \varepsilon} = \mathbb{E}\left(\left\| \frac{\tilde{Z}_{n,i} \tilde{\gamma}'_n \tilde{Z}_T - \tilde{Z}_T \tilde{\gamma}'_n \tilde{Z}_{n,i}}{(\tilde{Z}_{n,i} \tilde{\gamma}'_n)^2} \right\|^2 \|Y_{n,i}\|^2 \|(\hat{\gamma}'_n - \gamma_o)\|^2 \mathbf{1}_{\|\hat{\gamma}'_n - \gamma_o\| < \varepsilon}\right) < M^2 \varepsilon^2 < \eta.$$

Evaluating the error on the set  $\{|\hat{\gamma}_n - \gamma_o| < \varepsilon\}$ , we have

$$\begin{aligned} \mathbb{E}(\hat{Y}_T^{ScR} - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} &= \mathbb{E}(\hat{Y}_T^{ScR} - \hat{Y}_T^C)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} + \mathbb{E}(\hat{Y}_T^C - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} \\ &\quad - 2\mathbb{E}(\hat{Y}_T^{ScR} - \hat{Y}_T^C)(\hat{Y}_T^C - Y_T) \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} \end{aligned}$$

Moreover

$$\mathbb{E}(\hat{Y}_T^C - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} \leq \mathbb{E}(\hat{Y}_T^C - Y_T)^2 = \text{Var}(\hat{Y}_T^C - Y_T) = \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o}.$$

With the same argument as in theorem 3.3.4, when  $n$  is large enough

$$\mathbb{E}(\hat{Y}_T^C - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} = \mathbb{E}(\hat{Y}_T^C - Y_T)^2 - \mathbb{E}(\hat{Y}_T^C - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| \geq \varepsilon} > \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o} - \eta.$$

Using a similar argument as above, we can prove  $\forall \eta > 0, \exists \varepsilon > 0$  and  $n$  large enough such that

$$\|\mathbb{E}(\hat{Y}_T^{ScR} - \hat{Y}_T^C)(\hat{Y}_T^C - Y_T) \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon}\| < \eta$$

In other words

$$-3\eta + \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o} < \mathbb{E}(\hat{Y}_T^{ScR} - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} < 2\eta + \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o}.$$

Note that  $\mathbb{P}(\|\hat{\gamma}_n - \gamma_o\| < \varepsilon) \rightarrow 1$  as  $n \rightarrow \infty$  and the theorem holds.

Summing up all the target contained in the source  $S_{n,i}$  with the assumption that the number of targets nested within one source is bounded, we have

$$\mathbb{E}_{n,i}^{ScR} = \sum_{T \subset S_{n,i}} \mathbb{E}(\hat{Y}_T^{ScR} - Y_T)^2 \approx \sum_{T \subset S_{n,i}} \left[ \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o} \right] = \tilde{Z}_{n,i} \gamma_o - \sum_T \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o}.$$

This approximation yields

$$\text{Re}_{n,i}^{ScR} = \frac{\sqrt{\sum_{T \subset S_{n,i}} \mathbb{E}(\hat{Y}_T^{ScR} - Y_T)^2}}{\mathbb{E}(Y_{S_{n,i}})} \approx \frac{1}{\sqrt{\tilde{Z}_{n,i} \gamma_o}} \sqrt{\left[ 1 - \frac{\sum_{T \subset S_{n,i}} (\tilde{Z}_T \gamma_o)^2}{(\tilde{Z}_{n,i} \gamma_o)^2} \right]}.$$

# Bibliography

- Almquist, Z. W. et al. (2010). Us census spatial and demographic data in r: the uscensus2000 suite of packages. *J Stat Softw*, 37:1–31.
- Bajat, B., Krunić, N., and Kilibarda, M. (2011). Dasymeric mapping of spatial distribution of population in timok region. In *Proceedings of international scientific conference professional practice and education in geodesy and related fields, Kladovo, Serbia*, pages 30–34.
- Bloom, L., Pedler, P., and Wragg, G. (1996). Implementation of enhanced areal interpolation using mapinfo. *Computers & Geosciences*, 22(5):459–466.
- Calcagno, V., de Mazancourt, C., et al. (2010). glmulti: an r package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software*, 34(12):1–29.
- Chakir, R. (2009). Spatial downscaling of agricultural land-use data: an econometric approach using cross entropy. *Land Economics*, 85(2):238–251.
- Do, V. H., Thomas-Agnan, C., and Vanhems, A. (2015). Spatial reallocation of areal data - another look at basic methods. *Revue d'Économie Régionale et Urbaine*, pages 27–58.
- Dyn, N., Dyn, N., and Wong, W.-H. (1979). Comment. *Journal of the American Statistical Association*, 74(367):530–535.
- Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368.
- Fisher, P. F. and Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping\*. *The Professional Geographer*, 48(3):299–309.
- Flowerdew, R. and Green, M. (1993). Developments in areal interpolation methods and gis. In *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, pages 73–84. Springer.
- Flowerdew, R., Green, M., and Kehris, E. (1991). Using areal interpolation methods in geographic information systems. *Papers in regional science*, 70(3):303–315.
- Goodchild, M. F., Anselin, L., and Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, 25(3):383–397.
- Goodchild, M. F., Lam, N. S. N., and of Western Ontario. Dept. of Geography, U. (1980). *Areal interpolation: a variant of the traditional spatial problem*. London, Ont.: Department of Geography, University of Western Ontario.



- Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.
- 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 Economic Commission for Europe*, 17(2):157–188.
- Gregory, I. N. (2002). The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, environment and urban systems*, 26(4):293–314.
- Kelsall, J. and Wakefield, J. (2002). Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, 97(459):692–701.
- Kyriakidis, P. C. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3):259–289.
- Langford, M. (2007). Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems*, 31(1):19–32.
- Li, T., Pullar, D., Corcoran, J., and Stimson, R. (2007). A comparison of spatial disaggregation techniques as applied to population estimation for south east queensland (seq), australia.
- Liu, X., 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*, 22(4):431–447.
- Martin, D. (1989). Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, pages 90–97.
- Martin, D. and Bracken, I. (1991). Techniques for modelling population-related raster databases. *Environment and Planning A*, 23(7):1069–1075.
- Mennis, J. and Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, 33(3):179–194.
- 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*, pages 111–130.
- 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*, 95(451):877–887.
- Murakami, D. and Tsutsumi, M. (2011). A new areal interpolation method based on spatial statistics. *Procedia-Social and Behavioral Sciences*, 21:230–239.
- Rase, W.-D. (2001). Volume-preserving interpolation of a smooth surface from polygon-related data. *Journal of Geographical Systems*, 3(2):199–213.
- Reibel, M. and Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review*, 26(5-6):619–633.
- Reibel, M. and Bufalino, M. E. (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37(1):127–139.

- Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., and Ripley, M. B. (2015). Package ‘mass’.
- Sadahiro, Y. (1999). Accuracy of areal interpolation: A comparison of alternative methods. *Journal of Geographical Systems*, 1(4):323–346.
- Sadahiro, Y. (2000). Accuracy of count data estimated by the point-in-polygon method. *Geographical Analysis*, 32(1):64–89.
- Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530.
- Vignes, C., Rimbou, S., Ruiz-Gazen, A., and Thomas-Agnan, C. (2013). Fiches méthodologiques, méthodes statistiques d’allocation spatiale: interpolation de données surfaciques.
- Voss, P. R., Long, D. D., and Hammer, R. B. (1999). *When census geography doesn’t work: Using ancillary information to improve the spatial interpolation of demographic data*. Center for Demography and Ecology, University of Madison-Wisconsin.
- Wright, J. K. (1936). A method of mapping densities of population: With cape cod as an example. *Geographical Review*, pages 103–110.
- Yoo, E.-H., Kyriakidis, P. C., and Tobler, W. (2010). Reconstructing population density surfaces from areal data: A comparison of tobler’s pycnophylactic interpolation method and area-to-point kriging. *Geographical Analysis*, 42(1):78–98.
- 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*, 21(3):245–258.