

ACCURACY OF AREAL INTERPOLATION METHODS FOR COUNT DATA

DO Van Huyen ¹ & Christine THOMAS-AGNAN ² & Anne VANHEMS ³

¹ *huyendvmath@gmail.com*

² *Christine.Thomas@tse-fr.eu*

³ *a.vanhems@tbs-education.fr*

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. A particular case of that problem is the re-allocation to a single target partition which is a regular grid. 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 stochastic model based on Poisson point patterns to study the statistical accuracy of these techniques for regular grid targets in the case of count data. The error depends on the nature of the target variable and its correlation with the auxiliary variable. For simplicity, we restrict attention to proportional weighting schemes and Poisson regression based methods. Our conclusion is that there is no technique which always dominates.

Keywords. Areal interpolation, spatial disaggregation, pycnophylactic property, spatial misalignment, accuracy.

1 Introduction

The analysis of socio-economic data often involves the integration of various spatial data sources. Those data are often independently collected by a variety of offices and for different purposes. The zonal set systems used by distinct offices are rarely compatible and this leads to many difficulties. The problem of merging data bases on different spatial supports is called the areal interpolation or basis change problem (Goodchild and Lam 1980). In France, the need for official statistics at a more and more refined territorial level has been recognized by INSEE. In Europe, one of the objectives of the EU directive "INSPIRE", for INfrastructure for SPatial InfoRmation, is to harmonize quality geographic information to support the formulation and evaluation of public policies and activities which directly or indirectly impact the environment. Many statistical methods are proposed in the literature to handle this problem (dasymetric methods, regression

methods, smoothing techniques) and the reader is referred to Do et al. (2014) for a recent review of the simplest ones. The problem of their relative accuracy is most often treated at an empirical level (see for example Reibel and Bufalino (2005), Mennis (2006), Flowerdew and Green (1992), Flowerdew et al. (1991), Reibel and Agrawal (2007), Gregory (2002)). At the theoretical level, only few articles address this problem (Sadahiro (1999, 2000)) and this is the objective of this work.

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 section 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 2, 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 two sets of simulated data that we use later. In section 3, we recall the formulas for the dasymetric and Poisson regression areal interpolation. Finally, in section 4, we compare the relative accuracy of areal weighting and dasymetric methods with finite distance results whereas in section 5, 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 2. All proofs are in the appendix.

2 Count data and Poisson point pattern model

The variable of interest Y 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 Ω .

In the general area-to-area reallocation problem, the original data for the target variable is available for a set of source zones S_s ($s = 1, \dots, n_S$) and has to be transferred to an independent set of target zones T_t ($t = 1, \dots, n_T$). 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.

Overlaps between the two sets are called intersection zones and denoted by A_{st} for the intersection between the source S_s and the target T_t . For simplicity, $Y_{A_{st}}$ will be denoted by Y_{st} . Many methods involve the areas of different subregions (sources, targets or other).

We will denote by $|A|$ the area of any subregion A .

Most of 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 and Lam (1980), Langford (2005), 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) and Murakami (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. 2014 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 pattern 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 Ω , the extensive property is clearly satisfied

$$Y_\Omega = \sum_{i=1}^k Y_{\Omega_i}$$

With the proposed Poisson point pattern 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 λ_{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 pattern 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 Hulgren (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 (2011) utilizes 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 end of section 3). Therefore it corresponds to another underlying point process Z_X with intensity λ_{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 the two underlying point processes intensity functions

$$\lambda_{Z_Y}(s) = \alpha + \beta\lambda_{Z_X}(s), \quad (1)$$

where s is location. Therefore, 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 given by

$$Y_A \sim \mathcal{P}(\alpha|A| + \beta x_A) \quad (2)$$

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 βx_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 (1).

In sections 4.2 and 4.3, we introduce two sub-models of model (2) depending on the intensity function λ_{Z_Y} . We consider the case of a constant intensity (homogeneous model) and the case of a piecewise constant intensity (piecewise homogeneous model).

3 Methods

3.1 Prediction techniques

Do et al.(2014) classify areal interpolation methods into three groups: smoothing, dasymetric and regression based methods. We discard smoothing since it is concerned with continuous target variables which is not adapted to our count data model. We therefore focus here on the remaining two groups: dasymetric and regression based methods.

Dasymetric is a class of methods using a weighting scheme to allocate the original data to the intersections and then applying an aggregation step to get to target level. The simplest method in the dasymetric class is the areal weighting interpolation which uses area as weighting scheme: the data is allocated to the targets based on the assumption that the target variable is homogeneous at source level:

$$\hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \hat{Y}_{st} = \sum_{s:s \cap t \neq \emptyset} \frac{|A_{st}|}{|S_s|} Y_s.$$

Note that areal interpolation does not use any auxiliary information other than area which is usually available.

The general dasymetric method is supposed to improve upon the areal weighting interpolation method when an additional variable is known to be linked to the target variable leading to alternative weighting schemes. Voss et al. (1999) use road segment length and the number of road nodes for allocating demographic characteristics. Population, which is collected at fine levels in general, is often used as an auxiliary information for other variables like in Gregory (2002) or Mugglin and Carlin (1998). Instead of homogeneity, the dasymetric method with auxiliary information X assumes that the target variable is proportional to the auxiliary variable at intersection level.

$$\hat{Y}_t = \sum_{s:s \cap t \neq \emptyset} \hat{Y}_{st} = \sum_{s:s \cap t \neq \emptyset} \frac{X_{st}}{X_s} Y_s$$

where $X_s = \sum_t X_{st}$. This entails that X has to be known at intersection level, which is quite restrictive.

Concerning the regression based methods, there are several types of regression based methods also involving auxiliary information (see Do et al, 2014). Given the nature of the target variable in our model (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 (2), a Poisson regression with identity link is performed at source level yielding estimators $\hat{\alpha}, \hat{\beta}$ for the parameters α and β .

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

$$\hat{Y}_{st}^{REG} = \hat{\alpha}|A_{st}| + \hat{\beta}X_{st} \tag{3}$$

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.

One property often quoted concerning these methods is the pycnophylactic property. This property requires the preservation of the initial data in the following sense at some geographical level: at source level for example, it means that the predicted value for source S_s obtained by aggregating the predicted values on intersections with S_s should coincide with the observed value on S_s . The enforcement of this property will allow us to introduce an improved version of the basic Poisson regression method.

3.2 Prediction error criteria

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 2006,...) 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 5. We recall that we assume all target zones are nested within source zones.

In section 4, 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

$$\text{Er}_S^{MET} = \sum_{t \subset S} \text{Er}_t^{MET} = \sum_{t \subset S} \mathbb{E}(\hat{Y}_t^{MET} - Y_t)^2 \quad (4)$$

and the overall regional error is

$$\text{Er}^{MET} = \sum_S \sum_{t \in S} \mathbb{E}(\hat{Y}_t^{MET} - Y_t)^2 \quad (5)$$

In section 5, we use mean square error at target level

$$\text{Er}_t^{MET} = \mathbb{E}(\hat{Y}_t^{MET} - Y_t)^2 \quad (6)$$

to compare the dasymetric and Poisson regression methods.

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

$$\text{Re}_S^{MET} = \frac{\sqrt{\text{Er}_S^{MET}}}{\mathbb{E}(Y_S)} \quad (7)$$

where Re_S^{MET} is the relative error of method MET at source level for source S with method MET .

4 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 4.1, with two particular cases in 4.2 and 4.3, and a toy example in 4.4.

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

$$\hat{Y}_T^{DAW} = \frac{|T|}{|S|} Y_S \quad (8)$$

and

$$\hat{Y}_T^{DAX} = \frac{x_T}{x_S} Y_S \quad (9)$$

4.1 General auxiliary information model

Lemma 4.1 gives the expression of the prediction bias and variance in model AIM for areal weighting interpolation and dasymetric interpolation at target level.

Lemma 4.1. *In model AIM, the prediction biases and variances of areal weighting interpolation and dasymetric methods are given by*

$$\mathbb{E}(\hat{Y}_T^{DAW} - Y_T) = \beta x_S \left(\frac{|T|}{|S|} - \frac{x_T}{x_S} \right) \quad (10)$$

$$\mathbb{E}(\hat{Y}_T^{DAX} - Y_T) = \alpha |S| \left(\frac{x_T}{x_S} - \frac{|T|}{|S|} \right) \quad (11)$$

$$Var(\hat{Y}_T^{DAW} - Y_T) = \beta x_S \left(\frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_T \left(1 - \frac{x_T}{x_S} \right) + \alpha |T| \left(1 - \frac{|T|}{|S|} \right) \quad (12)$$

$$Var(\hat{Y}_T^{DAX} - Y_T) = \alpha |S| \left(\frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_T \left(1 - \frac{x_T}{x_S} \right) + \alpha |T| \left(1 - \frac{|T|}{|S|} \right) \quad (13)$$

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 $\left| \frac{x_S}{|S|} - \frac{x_T}{|T|} \right|$ 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 βx_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 4.2 reports the expression of the prediction variances in model AIM for areal weighting interpolation and dasymetric interpolation at source level.

Lemma 4.2. *In model AIM, the variances of areal weighting and dasymetric methods at the source level are*

$$Var_S^{DAW} = \beta x_S \sum_T \left(\frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_S \left(1 - \sum_T \frac{x_T^2}{x_S^2} \right) + \alpha |S| \left(1 - \sum_T \frac{|T|^2}{|S|^2} \right) \quad (14)$$

$$Var_S^{DAX} = \alpha |S| \sum_T \left(\frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_S \left(1 - \sum_T \frac{x_T^2}{x_S^2} \right) + \alpha |S| \left(1 - \sum_T \frac{|T|^2}{|S|^2} \right) \quad (15)$$

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

$$Var_S^{DAW} = \left(1 - \frac{1}{n_T} \right) (\alpha |S| + \beta x_S)$$

$$Var_S^{DAX} = \left(1 - \frac{1}{n_T} \right) (\alpha |S| + \beta x_S) + \sum_T \left(\frac{x_T^2}{x_S^2} - \frac{1}{n_T} \right) (\alpha |S| - \beta x_S)$$

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 a relative contribution of the corresponding effect to the overall mean at the geographical level of a subregion A :

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

$I_A(X)$ 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 4.1 and 4.2 the expression of the absolute and relative errors of the two methods at source level as a function of the relative contributions terms.

Theorem 4.3.

$$Er_S^{DAW} = I_S(X)^2 \mathbb{E}(Y_S)^2 D + I_S(X) \mathbb{E}(Y_S) (D + B) + I_S(|\cdot|) \mathbb{E}(Y_S) C \quad (16)$$

$$\Rightarrow (Re_S^{DAW})^2 = I_S(X)^2 D + \frac{1}{\mathbb{E}(Y_S)} [I_S(X)(D + B - C) + C] \quad (17)$$

$$(Re_S^{DAX})^2 = I_S(|\cdot|)^2 D + \frac{1}{\mathbb{E}(Y_S)} [I_S(|\cdot|)(D - B + C) + B] \quad (18)$$

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 the geometry of the problem and the auxiliary information, whereas the relative contribution terms and $\mathbb{E}(Y_S)$ depend on the coefficients α and β . 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

$$(Re_S^{DAW})^2 - (Re_S^{DAX})^2 = -D * \Delta_S \left(1 + \frac{1}{\mathbb{E}(Y_S)} \right) \quad (19)$$

which turns out to be clearly proportional to the imbalance term Δ_S . Similarly, one can approximate the ratio of the two relative errors when $\mathbb{E}(Y_S)$ is large on the target $A = T$ and on the source $A = S$ by

$$\frac{Re_A^{DAW}}{Re_A^{DAX}} \approx \frac{I_A(X)}{I_A(|\cdot|)} \quad (20)$$

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

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

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

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 the 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 Δ_S clearly determines which method is best.

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

$$\hat{Y}_T^C(w) = w\hat{Y}_T^{DAW} + (1-w)\hat{Y}_T^{DAX}, \quad (21)$$

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 4.5. *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 (21) is*

$$\hat{Y}_T^C = \hat{Y}_T^C(w^*) = \frac{\alpha|T| + \beta x_T}{\alpha|S| + \beta x_S} Y_S \quad (22)$$

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

$$Er_T^C = \frac{\lambda_T(\lambda_S - \lambda_T)}{\lambda_S} \quad (23)$$

$$(Re_S^C)^2 = \frac{1}{4\mathbb{E}(Y_S)} [\Delta_S^2 D + 2\Delta_S(C - B) + D + 2B + 2C] \quad (24)$$

Moreover, this predictor coincides with the best linear unbiased predictor in model AIM.

Because $\frac{\lambda_T(\lambda_S - \lambda_T)}{\lambda_S} = Var(\hat{Y}_T^{DAX} - Y_T) - \lambda_S(\frac{x_T}{x_S} - \frac{\lambda_T}{\lambda_S})^2 = Var(\hat{Y}_T^{DAW} - Y_T) - \lambda_S(\frac{|T|}{|S|} - \frac{\lambda_T}{\lambda_S})^2$, the prediction error of the best predictor is smaller than the variances of the other two methods and the distance is the more important that the auxiliary information is further from homogeneity. Of course, \hat{Y}_T^C is not a feasible predictor since it depends on the unknown coefficients α and β of model AIM but we will use it as a benchmark tool on the one hand and we will relate it later on to the regression predictor. If we look at the error at the level of source S , we have that $Er_S^C = \lambda_S - \sum_T \frac{\lambda_T^2}{\lambda_S} \leq \lambda_S - \frac{\lambda_S}{n_T(S)}$, 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 $\frac{\lambda_S}{n_T(S)}$. 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 Δ_S which minimizes the relative error of \hat{Y}_T^C

$$\Delta^* = \frac{B - C}{D} = \frac{\sum_T \left(\frac{|T|}{|S|}\right)^2 - \left(\frac{x_T}{x_S}\right)^2}{\sum_T \left(\frac{|T|}{|S|} - \frac{x_T}{x_S}\right)^2}$$

and thus derive a lower bound for the relative error

for a given geometry.

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.

4.2 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 (10), (12), (14) 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 (22) 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.

4.3 Piecewise homogeneous model

Suppose the source zone S comprises two homogeneous subzones C_1 and C_2 called control zones with intensities α_1 and α_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 4.1 and 4.2 give in this case

$$\begin{aligned} \mathbb{E}(\hat{Y}_T^{DAW} - Y_T)_{T:T \subset C_1} &= \frac{|T|}{|S|}(\alpha_2 - \alpha_1)|C_2| \\ \mathbb{E}(\hat{Y}_T^{DAW} - Y_T)_{T:T \subset C_2} &= \frac{|T|}{|S|}(\alpha_1 - \alpha_2)|C_1| \\ Var_S^{DAW} &= \alpha_1|C_1|(1 - \sum_{T:T \subset C_1} \frac{|T|^2}{|S|^2}) + \alpha_2|C_2|(1 - \sum_{T:T \subset C_2} \frac{|T|^2}{|S|^2}) \\ Er_S^{DAW} &= Var_S^{DAW} + \sum_{T:T \subset C_1} \frac{|T|^2}{|S|^2}(\alpha_2 - \alpha_1)^2|C_2|^2 + \sum_{T:T \subset C_2} \frac{|T|^2}{|S|^2}(\alpha_1 - \alpha_2)^2|C_1|^2 \end{aligned}$$

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.

4.4 Toy example

In order to illustrate our findings, we use a simulated toy example. We intentionally drop the assumption that targets are nested in sources which was made for mathematical convenience and this will allow us to test the robustness of the results with respect to that assumption. On a square grid with 25 cells, we design three sources and seven targets as unions of cells. On Figure 1, we see the design of sources and targets together with the cell counts for two target variables Y_1 and Y_2 and one auxiliary variable X (for one particular draw). To generate X , we simulate a Poisson point process with an inhomogeneous intensity. We then recover the counts at the cell level to get the auxiliary information. The two target variables are then generated according to their relationship with the auxiliary variable (model (1)) and the source values are obtained by aggregation of cells. The true value of two target variables is also shown at target level for accuracy comparison for one particular draw. For Y_1 , we use the set of parameters $\alpha = 80$ and $\beta = 1$ and for Y_2 we use $\alpha = 0$ and $\beta = 1$ so that the area has a strong impact on Y_1 and that Y_2 is only driven by X . Conditionally upon one draw of X (for which we observe 1011 points), we draw 1000 repetitions of Y_1 and Y_2 and present the relative error and the error at target level on Figure 2.

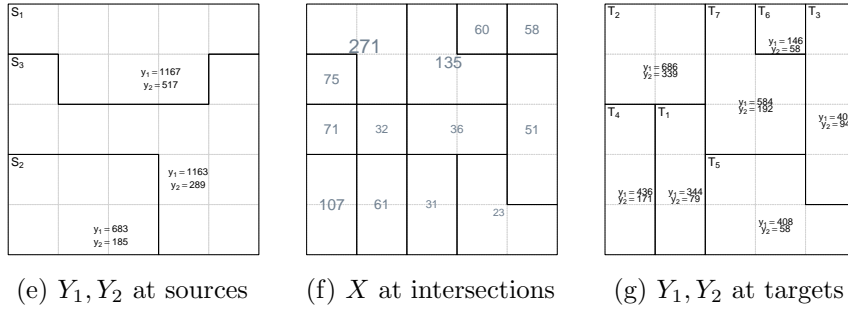
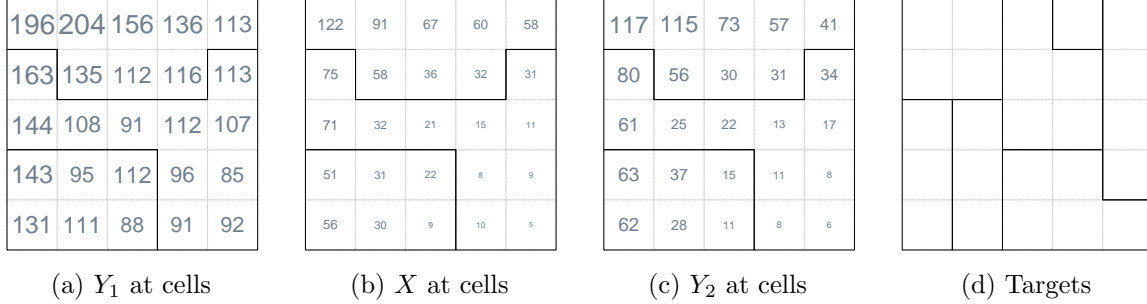


Figure 1: Toy example: Data at source, target, intersection zones.

The accuracy criterion is an average of the error over all 1000 draws. We present the relative error and the error at target level on Figure 2.

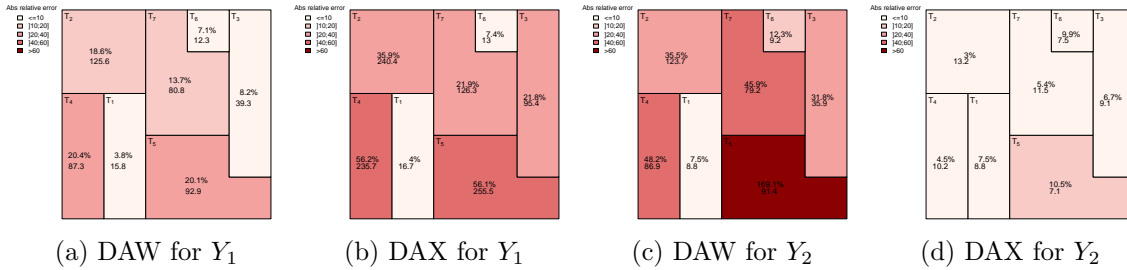


Figure 2: Toy example. Comparison of areal weighting interpolation and dasymetric method.

Figure 2 is in agreement with the theoretical results: areal weighting interpolation is better than dasymetric for Y_1 for which the areal effect dominates and worse for Y_2 for which the auxiliary information effect driven by X dominates. We can see that Y_2 is less homogeneous than Y_1 on Figure 1 (a) and (c): at the right-bottom target zone T_5 has

Table 1: **Square root of overall regional errors**

Methods	Y_1	Y_2
Areal weighting interpolation	201	197
Dasymetric	452	26
Regression	55	33

very small Y_2 counts. The dasymetric predictor is therefore very good for Y_2 (Figure 2 (d)). To be more precise, we compute the ratio of the two average errors at target level for the two methods and it shows that areal weighting is best for Y_1 with a ratio of square root of errors of $\frac{201}{452} = 44\%$, whereas dasymetric is best for Y_2 with ratio of square root of errors of $\frac{26}{197} = 13\%$. Table 1 reports the square root of the overall regional error (from formula (5)) for the three methods: areal weighting, dasymetric and Poisson regression. For Y_1 , the regression method is best, for Y_2 dasymetric is best because the impact of X is strong (almost no areal effect).

For the practitioner, an important question is to be able to guess which method will perform better in a given situation. We might believe that a good correlation between Y and X is a sign that dasymetric based on X will perform better than areal weighting. However in our case the correlation between Y_1 and X is 0.94 and the correlation between Y_2 and X is 0.98 which shows that this is a bad idea to rely on correlation. We could look at a measure of homogeneity to predict that areal weighting is the best method: in our case, the Gini coefficient of Y_1 is 0.14 and of Y_2 is 0.40 which goes in that direction. As we have seen in Theorem 4.4, the sign of the imbalance index at source level Δ_S determines which method is best (see Table 2).

To be more precise, let us examine the results of Table 3 in comparison with Theorem 4.4. Theorem 4.4 shows that the difference of errors at target level is influenced by three factors: the mean number of points of the source, the imbalance of the source and the inhomogeneity of the auxiliary information of the given target. Let us look at each impact. For the influence of the inhomogeneity of the auxiliary variable, we compare targets of a source S_1 for example. The first two impacts are constant (1164 points and 0.10 imbalance), and we see that the more homogenous the auxiliary information is (in increasing order $A_{16}, A_{13}, A_{17}, A_{12}$), the more distant the two methods are (18, 20, 1728, 2529 respectively). To examine the impact of the imbalance, we consider intersection zones A_{17} and A_{33} which are nested within two sources with similar number of points (S_1 with 1164 points and S_3 with 1168 points respectively). Even though their inhomogeneity are not very different (0.12 vs 0.10), the difference of the errors on A_{33} is 3.7 times larger than on A_{17} . This fact is explained by the distance between the two imbalances of S_1 and S_3 (0.10 vs 0.51). The last but not least effect is the number of points of the source zones. The comparison between A_{24} and A_{35} shows its impact: they have similar inhomogeneities (0.20 vs 0.19), not too different imbalances (0.41 vs 0.51), but the difference in errors for

Table 2: **Imbalance index at source level**

Source zones	S_1	S_2	S_3
Y_1	0.10	0.41	0.51
Y_2	-1	-1	-1

A_{35} (25707) is three times larger than the difference in errors for A_{24} (7938) and this is linked to the discrepancy in the mean numbers of points (1168 vs 679). As Table 3 shows, the combination of the three impacts is very complex. In other words, choosing between the areal weighting interpolation and the dasymmetric is not an easy problem.

Sources	S_1				S_2				S_3				
$\mathbb{E}(Y_1)$	1164				679				1168				
Imbalance	0.10				0.41				0.51				
Intersections	A_{12}	A_{13}	A_{16}	A_{17}	A_{21}	A_{24}	A_{25}	A_{31}	A_{32}	A_{33}	A_{34}	A_{35}	A_{37}
$\left \frac{ A }{ S } - \frac{X_A}{X_S} \right $	0.14	0.01	0.01	0.12	0.03	0.20	0.18	0.02	0.17	0.10	0.06	0.19	0.09
$\text{Er}^{DAX} - \text{Er}^{DAW}$	2529	20	18	1728	135	7938	5999	199	19858	6435	16817	25707	2240
$\text{Re}^{DAX} / \text{Re}^{DAW}$	1.5	1.3	1.1	1.4	1.8	5.4	5.4	3.1	8.8	7.5	8.7	8.8	6.0

Table 3: **Errors at intersection level for Y_1**

5 Relative accuracy of the other methods: asymptotic assessment

Let us now try to extend the comparison to the Poisson regression method. This cannot be done anymore by finite distance methods and so we introduce an asymptotic framework. Model (2) yields at source level

$$Y_s \sim \mathcal{P}(\alpha|S_s| + \beta x_s) \quad (25)$$

where $x_s = \sum_{t:t \cap s \neq \emptyset} x_{st}$. Besides the Poisson regression predictor defined by (3), inspired by Theorem 4.5, we propose a new predictor called scaled Poisson regression predictor defined as follows

$$\hat{Y}_{st}^{ScR} = \frac{\hat{\alpha}|A_{st}| + \hat{\beta}x_{st}}{\hat{\alpha}|S_s| + \hat{\beta}x_s} Y_S, \quad (26)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimators of α and β obtained through the Poisson regression at source level. Note that if model (1) 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 dasymetric 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 5.1, 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 5.2. Section 5.3 illustrates these results on a toy example.

5.1 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 Ω 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 Ω 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 $|\widetilde{A}| = k_n|A|, \tilde{x}_A = k_nx_A$.

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 $\lambda_A =$

$\gamma'Z_A, \tilde{Z}_A = \begin{pmatrix} |\widetilde{A}| \\ \tilde{x}_A \end{pmatrix} = k_nZ_A$ and $Y_A \sim \mathcal{P}(k_n\lambda_A)$. The true value of the parameter γ 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

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

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. $F_n^{1/2} F_n^{1/2} = F_n$.

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) $\{\tilde{Z}_{n,i}\} \subset \mathcal{Z} \forall n, i$ with \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.

$$c_1 < \|\tilde{Z}_{n,i}\| < c_2 \quad (27)$$

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

(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.

5.2 Predictors

In this section, we consider the asymptotic properties of the following two predictors: the regression predictor (3) and the scaled regression predictor (26). We prove that the scaled regression predictor is asymptotically as accurate as the unfeasible composite predictor. 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. It shows that, whereas it is satisfied by the scaled Poisson regression predictor, it is not satisfied at source level by the ordinary Poisson regression predictor but only at region level.

Proposition 5.2. *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.*

To prove proposition 5.2, we need the following asymptotic normality result for the target variables

$$\frac{Y_{ni} - \gamma'_o \tilde{Z}_{n,i}}{\sqrt{\gamma'_o \tilde{Z}_{n,i}}} \rightarrow_d \mathcal{N}(0, 1) \quad (28)$$

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

Theorem 5.3. *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 5.4. *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 relative error at source level can be approximated by

$$Re_{n,i}^{REG} \approx \frac{1}{\sqrt{\mathbb{E}(Y_{n,i})}} = \frac{1}{\sqrt{\alpha_o \widetilde{|S_{n,i}|} + \beta_o \tilde{x}_{n,i}}} \quad (29)$$

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. Equation (29) 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:

$$[(Re_{n,i}^{REG})^2 - (Re_{n,i}^{DAW})^2] \approx -(1 + \Delta_{n,i})^2 \sum_T \left(\frac{|T|}{|S_{n,i}|} - \frac{x_T}{x_{n,i}} \right)^2 \quad (30)$$

$$[(Re_{n,i}^{REG})^2 - (Re_{n,i}^{DAX})^2] \approx -(1 - \Delta_{n,i})^2 \sum_T \left(\frac{|T|}{|S_{n,i}|} - \frac{x_T}{x_{n,i}} \right)^2 \quad (31)$$

$$(32)$$

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 $\left(\frac{|T|}{|S_{n,i}|} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right) = 0$ then the regression is less accurate than areal weighting and dasymetric asymptotically so that none of them is always dominant.

For areal weighting and dasymetric predictors, we have seen that if one method is better on one target, then it is also true on all targets contained on the same source zone. 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 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 composite predictor. The first lemma proves an asymptotic equivalence between the scaled regression predictor and the unfeasible composite predictor.

Lemma 5.5. *For any target T ,*

$$\hat{Y}_T^{ScR} - \hat{Y}_T^C \xrightarrow{p} 0 \quad (33)$$

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

Theorem 5.6. *For any $\eta > 0$, there exists a sequence of sets $\{Q_i\}_i : \mathbb{P}(Q_i) \rightarrow 1$ such that*

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

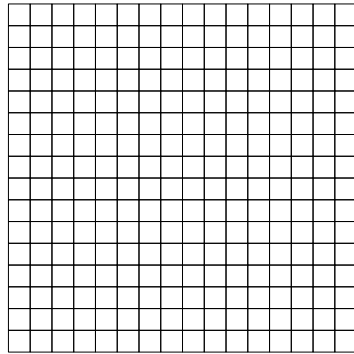
Since $\text{Er}_T^C = \tilde{Z}_T \gamma_o - \frac{(\tilde{Z}_T \gamma_o)^2}{\tilde{Z}_{n,i} \gamma_o}$, this theorem shows that the quadratic prediction error of the scaled regression predictor is asymptotically equivalent to the one of the composite. Consequently, the scaled regression method is the best among the areal weighting, the dasymetric and the regression predictors.

5.3 Accuracy: simulation assessment with a toy example

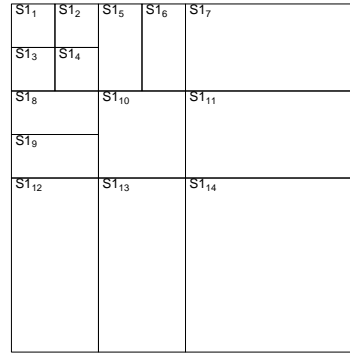
We devise a simple simulation to illustrate these results. On a square region Ω with 16×16 cells, we build three systems of sources with respectively 14, 7 and 4 sources (see Figure 5.3). We simulate two Poisson point patterns (our auxiliary information) with an expected overall number of points of 100,000: 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 (2). For each of the auxiliary variables, we choose four couples of coefficients α, β to study the effects of imbalance so that we get eight different target variables. Table 4 reports the minimum, maximum and average imbalance for each case and for each system of source zones.

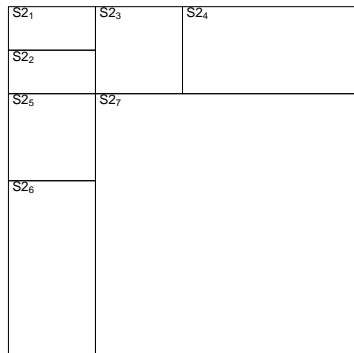
We then apply the four considered methods (areal weighting, dasymetric, Poisson regression and scaled Poisson regression) to transfer the data from each of the three systems of source zones to cell level. For each case, we generate the data 1000 times, and calculate prediction errors for each method and each iteration. Table 5 (respectively Table 6) reports the average absolute square root of prediction errors (respectively the average absolute square root of relative prediction errors). The two tables also present the mean of the target variables at region level $\mathbb{E}(Y_\Omega) = \alpha|\Omega| + \beta x_\Omega$ (because it appears in Theorem 5.4) and the theoretical composite prediction error as a benchmark (see Theorem 5.6).



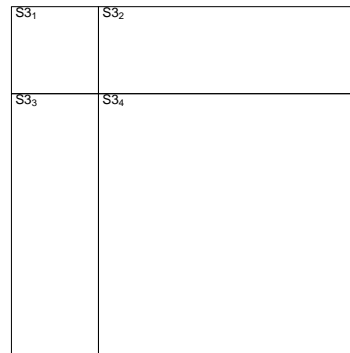
(a) Cells



(b) Sources 1



(c) Sources 2



(d) Sources 3

Figure 3: Spatial polygons

Table 5 shows that when β is fixed and α increases, resulting in an increase of the mean of the target variables at region level ($\alpha|\Omega| + \beta x_\Omega$), all errors get larger. For fixed coefficients α, β , the errors increase from the first set of sources to the third one, which is natural since the available information decreases from 14 observations for the first case, to 4 for the third.

In accordance with the toy example of section 4.4, the errors for X_2 are much smaller than the ones for X_1 for the areal weighting and the dasymetric methods due to the difference of homogeneity of the auxiliary variables. The more homogeneous the auxiliary variable is, the more accurate the areal weighting interpolation and dasymetric methods

Table 4: **Imbalances**

Case	Sources	Min	Mean	Max
$\alpha = 100$ $\beta = 1$ X_1	14 sources	-0.92	-0.40	0.96
	7 sources	-0.90	-0.59	0.64
	4 sources	-0.87	-0.44	0.64
$\alpha = 600$ $\beta = 1$ X_1	14 sources	-0.62	0.11	0.99
	7 sources	-0.53	-0.05	0.93
	4 sources	-0.43	0.07	0.93
$\alpha = 1000$ $\beta = 1$ X_1	14 sources	-0.43	0.29	1
	7 sources	-0.33	0.17	0.96
	4 sources	-0.21	0.26	0.96
$\alpha = 1000$ $\beta = 0.1$ X_1	14 sources	0.6	0.86	1
	7 sources	0.67	0.84	1
	4 sources	0.74	0.86	1

Case	Sources	Min	Mean	Max
$\alpha = 100$ $\beta = 1$ X_2	14 sources	-0.63	-0.6	-0.58
	7 sources	-0.61	-0.6	-0.59
	4 sources	-0.6	-0.6	-0.59
$\alpha = 600$ $\beta = 1$ X_2	14 sources	0.16	0.21	0.22
	7 sources	0.18	0.2	0.22
	4 sources	0.19	0.2	0.22
$\alpha = 1000$ $\beta = 1$ X_2	14 sources	0.39	0.43	0.45
	7 sources	0.41	0.43	0.44
	4 sources	0.42	0.43	0.44
$\alpha = 1000$ $\beta = 0.1$ X_2	14 sources	0.92	0.92	0.93
	7 sources	0.92	0.92	0.93
	4 sources	0.92	0.92	0.93

are. As we discussed earlier, if the auxiliary information is almost homogeneous, the regression might be less exact than the areal weighting. But Table 4 shows that the regression methods are still quite good for X_2 : in general, they are better than the two classical methods except in some particular cases. The errors of the regression and scaled regression methods are very comparable for X_1 and X_2 . Indeed, the prediction errors of the regression are very similar to the mean of the target variables Y and the accuracy of the scaled regression predictor is equivalent to the one of the composite predictor.

The effect of imbalance can be studied by looking at a change of α with fixed β . A larger α corresponds to a larger influence of the areal effect $\alpha|S|$ which is expected to lead to the domination of the areal weighting interpolation method (indeed we can observe this effect in the table for both auxiliary variables X_1 and X_2). The imbalance also affects the regression and scaled regression methods: if one of the two effects $\alpha|S|$ and βX_S is much larger than the other one, the corresponding errors seem to be further from their benchmarks (respectively the mean of Y and the composite prediction error): see for example the cases $\alpha = 1000, \beta = 0.1$. However this effect is not very large for regression and scaled regression. One factor which influences more these two methods is the homogeneity of the auxiliary variable: comparing the results for X_1 and X_2 illustrates this. For X_1 , the regression prediction errors are almost equal to their respective benchmarks and the amount of initial information does not seem to have a big influence (the errors are not monotonic from the first to the third set of sources). For X_2 , the accuracy increases with the number of source zones (the best being for the first one) and the errors of the regression method tend to the mean of Y . If we consider the particular case $\alpha = 1000, \beta = 0.1$ for X_2 , the areal impact is much stronger than the auxiliary information impact, and we see that the areal weighting interpolation is the best method, and that the scaled regression predictors

Table 5: Square root of prediction errors.

Methods	$\sqrt{\mathbb{E}(Y_\Omega)}$	Sources	DAW	DAX	REG	ScR	Composite
$\alpha = 100$ $\beta = 1$ X_1	354.7	14 sources	6580.5	1621.6	353.0	333.1	332.1
		7 sources	6962.7	2087.7	352.1	341.3	341.2
		4 sources	7215.7	2090.6	352.1	347.6	347.8
$\alpha = 600$ $\beta = 1$ X_1	503.8	14 sources	6589.6	9532.5	500.6	481.9	482.1
		7 sources	6971.3	12363.7	502.9	493.4	491.7
		4 sources	7225.3	12374.7	502.5	498.8	497.4
$\alpha = 1000$ $\beta = 1$ X_1	596.9	14 sources	6597.3	15878.2	594.0	574.5	574.2
		7 sources	6982.2	20595.2	595.6	586.5	584.6
		4 sources	7229.6	20614.5	594.4	590.8	590.3
$\alpha = 1000$ $\beta = 0.1$ X_1	515.8	14 sources	826.7	15875.6	513.6	500.7	500.9
		7 sources	861.9	20592.8	515.0	509.4	508.3
		4 sources	883.5	20612.4	514.5	512.3	511.6
$\alpha = 100$ $\beta = 1$ X_2	354.4	14 sources	458.8	353.5	356.6	348.3	344.5
		7 sources	469.0	358.6	357.7	354.2	349.5
		4 sources	474.2	361.3	359.6	358.2	351.6
$\alpha = 600$ $\beta = 1$ X_2	503.6	14 sources	573.9	677.4	504.9	492.9	489.6
		7 sources	587.3	690.0	508.1	503.1	496.6
		4 sources	591.2	697.7	509.1	507.2	499.6
$\alpha = 1000$ $\beta = 1$ X_2	596.7	14 sources	654.8	969.9	599.0	585.0	580.1
		7 sources	665.5	992.6	601.7	595.7	588.4
		4 sources	671.8	1002.6	602.2	600.0	592.0
$\alpha = 1000$ $\beta = 0.1$ X_2	515.8	14 sources	502.9	924.7	518.6	506.8	501.4
		7 sources	510.3	947.4	520.4	515.5	508.6
		4 sources	512.9	960.9	522.2	520.2	511.7

Methods	$\mathbb{E}(Y_\Omega)$	Sources	DAW	DAX	REG	ScR	Composite
$\alpha = 100$ $\beta = 1$ X_1	125847	14 sources	50.459	47.464	6.979	6.872	6.873
		7 sources	55.664	54.021	6.970	6.941	6.951
		4 sources	57.122	54.054	6.978	6.966	6.969
$\alpha = 600$ $\beta = 1$ X_1	253847	14 sources	20.458	59.087	3.492	3.422	3.421
		7 sources	21.865	73.067	3.495	3.471	3.466
		4 sources	22.557	73.208	3.490	3.480	3.481
$\alpha = 1000$ $\beta = 1$ X_1	356247	14 sources	14.859	62.210	2.813	2.754	2.757
		7 sources	15.827	77.405	2.823	2.801	2.794
		4 sources	16.324	77.589	2.820	2.811	2.808
$\alpha = 1000$ $\beta = 0.1$ X_1	266025	14 sources	4.297	72.322	3.093	3.018	3.020
		7 sources	4.433	89.662	3.100	3.069	3.064
		4 sources	4.508	90.028	3.098	3.086	3.082
$\alpha = 100$ $\beta = 1$ X_2	125608	14 sources	5.685	4.503	4.545	4.438	4.391
		7 sources	5.799	4.569	4.559	4.514	4.455
		4 sources	5.859	4.603	4.581	4.564	4.482
$\alpha = 600$ $\beta = 1$ X_2	253608	14 sources	3.574	4.121	3.186	3.108	3.088
		7 sources	3.655	4.196	3.205	3.174	3.134
		4 sources	3.679	4.236	3.211	3.199	3.153
$\alpha = 1000$ $\beta = 1$ X_2	356008	14 sources	2.918	4.084	2.692	2.627	2.606
		7 sources	2.966	4.172	2.704	2.677	2.645
		4 sources	2.993	4.211	2.706	2.696	2.661
$\alpha = 1000$ $\beta = 0.1$ X_2	266001	14 sources	3.022	5.133	3.118	3.045	3.015
		7 sources	3.068	5.246	3.129	3.099	3.059
		4 sources	3.084	5.320	3.140	3.128	3.078

Table 6: **Relative prediction errors (in percentages).**

can catch up the areal weighting interpolation when there are more source zones.

Table 6 contains the corresponding relative errors. For example, in the case $\alpha = 1000, \beta = 1$ for X_2 , we see that whatever the number of sources the relative error is around 2.6% to 2.7% for the scaled regression (very close to the benchmark given by the last column) whereas dasymetric is around 4% and areal weighting around 3%. Looking at the second column, we see that when the expected number of points increases, the relative prediction error tends to decrease which was naturally not the case for the prediction error itself.

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 5.3) 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 7 presents the results.

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 7: **Robusness of methods.**

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 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.

6 Conclusion

In this paper 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.

7 Appendix

7.1 Proofs

7.1.1 Proof of Lemma 4.1 and lemma 4.2

From (8), (9) and the properties of a Poisson point process we have

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

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

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

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

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

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

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

$$\begin{aligned}
Var_S^{DAX} &= \sum_T Var(\hat{Y}_T^{DAX} - Y_T) \\
&= \alpha|S| \sum_T \left(\frac{|T|}{|S|} - \frac{x_T}{x_S} \right)^2 + \beta x_S \left(1 - \sum_T \frac{x_T^2}{x_S^2} \right) + \alpha|S| \left(1 - \sum_T \frac{|T|^2}{|S|^2} \right)
\end{aligned}$$

7.1.2 Proof of Theorem 4.3

From Lemma 4.1 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

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

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

$$\frac{Er_T^{DAW}}{Er_T^{DAX}} \approx \frac{I_S(X)^2}{I_S(|\cdot|)^2}$$

and also

$$\frac{Re_T^{DAW}}{Re_T^{DAX}} \approx \frac{I_S(X)}{I_S(|\cdot|)}$$

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

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

Using the relationship $I_S(|\cdot|) + I_S(X) = 1$, the above results prove Theorem 4.3.

7.1.3 Proof of Theorem 4.4

Lemma 4.1 yields

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

7.1.4 Proof of Theorem 4.5

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

$$\begin{aligned}\hat{Y}_T^C &= w\hat{Y}_T^{DAW} + (1-w)\hat{Y}_T^{DAX} = [w\frac{|T|}{|S|} + (1-w)\frac{x_T}{x_S}]Y_S := uY_S \\ Bias_T^2 &= [\mathbb{E}(\hat{Y}_T^{DAW} - \hat{Y}_T^{DAX})]^2 = (u\lambda_S - \lambda_T)^2 \\ Var_T &= Var(uY_S - Y_T) = u^2\lambda_S + \lambda_T - 2u\lambda_T \\ Er_T &= u^2\lambda_S(\lambda_S + 1) - 2u\lambda_T(\lambda_S + 1) + \lambda_T^2 + \lambda_T \\ u^* &= argmin_u Er_T = \frac{\lambda_T}{\lambda_S} \\ \Leftrightarrow w^* \frac{|T|}{|S|} + (1-w^*)\frac{x_T}{x_S} &= \frac{\alpha|T| + \beta x_T}{\alpha|S| + \beta x_S} \\ \Leftrightarrow w^* &= \frac{\alpha|T|}{\alpha|S| + \beta x_S}\end{aligned}$$

Substituting the w^* in (21) we get the composite predictor (22).

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

$$\begin{aligned}Bias &= \mathbb{E}(\hat{Y}_T^C - Y_T) = 0 \\ Er_T^C &= Var(\hat{Y}_T^C - Y_T) = Var(\frac{\lambda_T}{\lambda_S}Y_S - Y_T) \\ &= \frac{\lambda_T^2}{\lambda_S^2}Var(Y_S) + Var(Y_T) - 2\frac{\lambda_T}{\lambda_S}Cov(Y_S, Y_T) \\ &= \frac{\lambda_T^2}{\lambda_S^2}\lambda_S + \lambda_T - 2\frac{\lambda_T}{\lambda_S}\lambda_T \\ &= \lambda_T - \frac{\lambda_T^2}{\lambda_S} \\ &= \frac{x_T^2}{x_S^2}\lambda_S + \lambda_T - 2\frac{x_T}{x_S}\lambda_T - \frac{x_T^2}{x_S^2}\lambda_S + 2\frac{x_T}{x_S}\lambda_T - \frac{\lambda_T^2}{\lambda_S} \\ &= \frac{x_T^2}{x_S^2}\lambda_S + \lambda_T - 2\frac{x_T}{x_S}\lambda_T - \lambda_S(\frac{x_T}{x_S} - \frac{\lambda_T}{\lambda_S})^2 \\ &= Var(\hat{Y}_T^{DAX} - Y_T) - \lambda_S(\frac{x_T}{x_S} - \frac{\lambda_T}{\lambda_S})^2 \\ &= Var(\hat{Y}_T^{DAW} - Y_T) - \lambda_S(\frac{|T|}{|S|} - \frac{\lambda_T}{\lambda_S})^2\end{aligned}$$

Since

$$Y_T|Y_S \sim Bi(Y_S, \frac{\mathbb{E}(Y_T)}{\mathbb{E}(Y_S)})$$

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 composite predictor is the best linear predictor.

7.1.5 Proof of Theorem 5.1

To prove the theorem, we will prove the following lemmas

Lemma 7.1. *Under conditions (C1) and (C2), the normed score function $F_n^{-1/2}s_n$ is asymptotically normal*

$$F_n^{-1/2}s_n \rightarrow_d \mathcal{N}(0, \mathbf{I}) \quad (34)$$

Lemma 7.2. *Under conditions (C1) and (C2), for all $\delta > 0$*

$$\max_{\gamma \in N_n(\delta)} \|V_n(\gamma) - \mathbf{I}\| \rightarrow_p 0 \quad (35)$$

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 7.1 is proved by using the Lindeberg-Feller theorem.

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

$$z_{n,i} = \tau' F_n^{-1/2} \frac{\tilde{Z}_{n,i}}{\gamma' \tilde{Z}_{n,i}} (y_{n,i} - \gamma' \tilde{Z}_{n,i}) \quad (36)$$

we have

$$\begin{aligned} \mathbb{E}(z_{n,i}) &= 0 \\ \sum_i \text{Var}(z_{n,i}) &= 1 \end{aligned}$$

We will show that the Lindeberg condition is satisfied, i.e. for any $\varepsilon > 0$

$$\sum_i \mathbb{E}(z_{n,i}^2 \mathbf{1}_{|z_{n,i}| > \varepsilon}) \rightarrow 0 \quad (37)$$

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

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

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. $\forall n > n_1$

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

hence

$$\sup_i \mathbb{E}((y_{n,i} - \gamma' \tilde{Z}_{n,i})^2 \mathbf{1}_{|y_{n,i} - \gamma' \tilde{Z}_{n,i}| > \frac{\varepsilon}{|a_{n,i}|}}) \rightarrow 0 \text{ as } n \rightarrow \infty$$

where the existence of K_2 is derived from condition (C1). This argument shows that the (37) holds. So does Lemma 7.1.

Proof of Lemma 7.2

Using the same notation in the proof of Lemma 7.1, τ fixed s.t. $\tau' \tau = 1$, let $b_{n,i} = \tau' F_n^{-1/2} \tilde{Z}_{n,i}$, the equation (35) can be rewritten as

$$\tau'(V_n(\gamma) - \mathbf{I})\tau = A_n + B_n + C_n \quad (38)$$

where

$$A_n = \sum_i b_{n,i}^2 \left(\frac{1}{(\gamma' \tilde{Z}_{n,i})^2} - \frac{1}{(\gamma'_o \tilde{Z}_{n,i})^2} \right) (y_{n,i} - \gamma' \tilde{Z}_{n,i}) \quad (39)$$

$$B_n = \sum_i b_{n,i}^2 \frac{1}{(\gamma'_o \tilde{Z}_{n,i})^2} (y_{n,i} - \gamma' \tilde{Z}_{n,i}) \quad (40)$$

$$C_n = \sum_i b_{n,i}^2 \left(\frac{1}{\gamma' \tilde{Z}_{n,i}} - \frac{1}{\gamma'_o \tilde{Z}_{n,i}} \right) \quad (41)$$

We will prove that the three terms converge in probability to 0 as n tends to ∞ . To prove (40), 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 7.2 holds.

7.1.6 Proof of Theorem 5.3

Let

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

This yields

$$\sum_j z_{nij} = Y_{n,i} - \gamma'_o \tilde{Z}_{n,i}$$

We have

$$\begin{aligned}\mathbb{E}(z_{nij}) &= 0 \\ \sum_j \text{Var}(z_{nij}) &= \gamma'_o \tilde{Z}_{n,i}\end{aligned}$$

We will prove that this array satisfies the Lindeberg-Feller condition, i.e. $\forall \delta > 0$

$$\sum_j \mathbb{E}(z_{nij}^2 \mathbf{1}_{|z_{nij}| > \delta}) \rightarrow 0, \text{ as } n \rightarrow \infty$$

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.

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

7.1.7 Proof of Proposition 5.2

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

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

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

$$\begin{aligned}
& \sum_{i=1}^n \frac{\tilde{Z}_{n,i}}{\hat{\gamma}' \tilde{Z}_{n,i}} y_{n,i} - \tilde{Z}_{n,i} = 0 \\
\Rightarrow & \sum_{i=1}^n \frac{\hat{\gamma}' \tilde{Z}_{n,i}}{\hat{\gamma}' \tilde{Z}_{n,i}} y_{n,i} - \hat{\gamma}' \tilde{Z}_{n,i} = 0 \\
\Leftrightarrow & \sum_{i=1}^n y_{n,i} - \hat{\gamma}' \tilde{Z}_{n,i} = 0 \\
\Leftrightarrow & \hat{\gamma}' \tilde{Z}_{n,i} = y_{n,i}
\end{aligned}$$

In other words, the regression predictor satisfies the pycnophylactic property on the region Ω .

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 5.1. The second term is different from 0, even asymptotically (Proposition 5.2).

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

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

This completes the proof of proposition 5.2.

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

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

7.1.8 Proof of Theorem 5.4

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

$$\mathbb{E}(\hat{\gamma}'_n \tilde{Z}_T - \gamma'_o \tilde{Z}_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} < \eta_1 \quad (42)$$

$$\|2\mathbb{E}(\hat{\gamma}'_n \tilde{Z}_T - \gamma'_o \tilde{Z}_T)(\gamma'_o \tilde{Z}_T - Y_T) \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon}\| < \eta_1 \quad (43)$$

As we proved in Proposition 5.2 $(\hat{\gamma}'_n \tilde{Z}_T - \gamma'_o \tilde{Z}_T) \rightarrow_p 0$, we have

$$\mathbb{P}(\|\hat{\gamma}_n - \gamma_o\| < \varepsilon) \rightarrow 1 \text{ as } n \rightarrow \infty$$

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 $n > n_1$

$$-\eta + \gamma'_o \tilde{Z}_T < \mathbb{E}(\hat{Y}_T^R - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} < \eta + \gamma'_o \tilde{Z}_T \quad (44)$$

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

Combining (42), (43), (44) we get Theorem 5.4.

7.1.9 Proof of equations (30)

We rewrite the error of the areal interpolation and dasymmetric for the asymptotic model. For a target $T \subset S_{n,i}$, from (23), (24), and Lemma 4.1 we have

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

A similar argument as in the proof of theorem 5.4 shows that, for any $\eta_1 > 0, \varepsilon > 0, \exists n_1$ s.t. $\forall n > n_1$

$$\begin{aligned}\mathrm{Er}_T^{DAW} - \eta_1 &< \mathbb{E}(\hat{Y}_T^{DAW} - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} < \mathrm{Er}_T^{DAW} \\ \mathrm{Er}_T^{DAX} - \eta_1 &< \mathbb{E}(\hat{Y}_T^{DAX} - Y_T)^2 \mathbf{1}_{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon} < \mathrm{Er}_T^{DAX}\end{aligned}$$

With ε chosen as in theorem 5.4, let $Q_i = \{\|\hat{\gamma}_n - \gamma_o\| < \varepsilon\}$, we have

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

for all $n > n_1$. Moreover,

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

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

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

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

If $\left(\frac{|\widetilde{T}|}{|\widetilde{S_{n,i}}|} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right) = 0$ then the regression is less accurate than areal weighting and dasy-metric 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 $\left(\frac{|\widetilde{T}|}{|\widetilde{S_{n,i}}|} - \frac{\tilde{x}_T}{\tilde{x}_{n,i}} \right) \neq 0$, we have

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

therefore,

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

Choosing η, η_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.

7.1.10 Proof of Lemma 5.5

Assume $T \in S_{n,i}$, the difference between the predictors of scaled regression and composite 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 γ_o .

From Theorem 5.1, property (28), conditions (C1), (C2), we have

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

In other words,

$$\hat{Y}_T^{ScR} - \hat{Y}_T^C \rightarrow_p 0$$

7.1.11 Proof of Theorem 5.6

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

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

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

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

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 5.4, when n is large enough

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

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.

References

- [1] Do V. H., Thomas-Agnan C. and Vanhems A., 2014, Spatial reallocation of areal data: a review, (to appear in RERU).
- [2] Flowerdew, R. and Green, M. (1992), Developments in areal interpolation methods and GIS, *The Annals of Regional Science*, 26, 67–78.
- [3] Flowerdew, R., Green, M and Kehris, E. (1991), Using areal interpolation methods in geographic information systems. *Papers in Regional Science*, Vol. 70, Issue 3, 303-315
- [4] Goodchild, M.F, Lam, N.S.-N. (1980), Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing*, 1, 297-312.
- [5] Goodchild, M.F, Anselin, L. and Deichman, U. (1993), A framework for the areal interpolation of socio-economic data, *Environment and Planning A*, 25, 383-397.

- [6] Gregory (2002), The accuracy of areal interpolation techniques : standardizing 19th and 20th century census data to allow long-term comparisons, *Computers, environments and urban systems* 26, 293-314.
- [7] Mennis, .J and Hultgren, T. (2006), Intelligent dasymetric mapping and its application to areal interpolation, *Cartography and Geographic Information Science*, vol. 33, n^o 3, pp. 179-194.
- [8] Mugglin, A.S. and Carlin B.P., (1998), Hierarchical modeling in geographic information systems: ppulation interpolation over incompatible zones, *Journal of Agricultural, Biological and Environmental Statistics*, vol. 3, n^o 2, pp. 111-130.
- [9] Murakami, D. and Tsutsumi, M., (2011), A New Areal Interpolation Technique Based on Spatial Econometrics, *Procedia-Social and Behavioral Sciences*, vol. 21, pp. 230-239.
- [10] Langford, M., (2007), Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps, *Computers, Environment and Urban Systems*, vol. 31, pp. 19-32.
- [11] Li, T., Pullar, D. V., Corcoran, J. and Stimson, R. J. (2007) A comparison of spatial disaggregation techniques as applied to population estimation for south east Queensland (SEQ), Australia. *Applied GIS*, vol.3 issue 9, pp. 1-16.
- [12] Ludwig Fahrmeir and Heinz Kaufmann (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *The Annals of Statistics*, Vol. 13, No. 1, 342-368.
- [13] Reibel, M. and Bufalino, M.E., (2005), Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems, *Environment and Planning A*, vol. 37, n^o 1, pp. 127-139.
- [14] Reibel, M. and Agrawal, A., (2007), Areal interpolation of population counts using pre-classified land cover data, *Population Research and Policy Review*, vol. 26, n^o 5-6, pp. 619-633.
- [15] Sadahiro, Y., (1999), Accuracy of areal interpolation: A comparison of alternative methods, *Journal of Geographical Systems*, vol. 1, Issue 4, pp. 323-346.
- [16] Sadahiro, Y., (2000), Accuracy of count data estimated by the point-in-polygon method, *Geographical Analysis*, vol. 32, Issue 1, pp. 64-89.
- [17] Voss, P.R., Long, D.L., and Hammer, R.B. (1999) When census geography doesn't work: using ancillary information to improve the spatial interpolation of demographic data, CDE working paper 99-26, Wisconsin, Madison.

- [18] Yew Yuan, Richard M. Smith and W. Fredrick Limp (1997) Remodelling census population with spatial information from landsat TM imagery, *Comput., Environ. and Urban Systems*, Vol. 21, No. 3/4, pp. 245-258.