TESTING AREAL INTERPOLATION METHODS WITH US CENSUS 2010 DATA

Huyen DO VAN[1], Christine THOMAS-AGNAN[2], Anne VANHEMS[3]

Abstract - *The areal interpolation problem is that of projecting a characteristic of interest on a partition of space called target partition from the knowledge of the same characteristic on a different partition, so called source partition, using some auxiliary information. The objective of this paper is to use a demographic database available in the R package 'US census 2010' (Almquist, 2010) in order to test several areal interpolation methods based on regression in the case of count related data. The fact that data is available at many different spatial scales in this database make this comparison study unique. Another innovative point of view is that we compare the extensive approach versus the intensive approach for a variable which is a ratio of counts. We also include the comparison with the scaled regression for the extensive case introduced in Do et al. (2015) and with a scaled regression for the intensive case proposed here. Finally we give some empirical guidelines for the choice of auxiliary information.*

Keywords - AREAL INTERPOLATION, SPATIAL DISAGGREGATION, PYCNOPHYLACTIC PROPERTY

Classification JEL - C83; C15; R15

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

[1] Toulouse School of Economics (GREMAQ) ; huyendvmath@gmail

[2] Toulouse School of Economics (GREMAQ) ; Christine.Thomas@tse-fr.eu

[3] Toulouse Business School and Toulouse School of Economics ; a.vanhems@tbs-education.fr

1. INTRODUCTION

The areal interpolation problem often arises in the analysis of socio-economic data involving the combination of several databases originating from different administrative sources. It is the problem of transferring a given variable called target variable known on an initial partition of space into source zones to another partition of the same space into target zones. Do et al. (2014) review these statistical spatial allocation rules concentrating on the simplest ones. Do et al. (2015) analyze the accuracy of these methods in the case of count data. For a set of random points in a region of space, a count variable is defined by the numbers of these points in given subdivisions of this region. Demographic data are of this nature since they are counts of different types of people in different subdivisions of space. A related type of variable is a density variable: number of points per areal unit associated to a set of random points in space. A count variable is said extensive when its value on a region is obtained by summing up its values on any partition into subregions. A density variable is said intensive in the sense that its 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. A more general case of count related variable that we will consider is a ratio of two count variables which turns out to be intensive.

The aim of this paper is to derive some empirical guidelines of application for several areal interpolation methods and confront the empirical evidence with some theoretical results of Do et al. (2015) using the demographic database ‘US census 2010’ (Almquist, 2010) available in an R package. We concentrate on count related data and focus particularly on the following points. The data is presented in section 2. We recall the main methods in section 3. In section 4.1 we give some directives for the selection of a good auxiliary variable when there is a choice. In section 4.2 we compare the accuracy of the regression methods. The regression methods developed for the extensive case are different from the ones adapted to the intensive case. Since it is easy to transform an extensive variable into a corresponding intensive one and reversely, we explore in section 4.3 the question of whether it is best to use the extensive-type or intensive-type regression method. Finally we investigate in section 4.4 the effect of the spatial scale.

2. DATA

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

Following Almquist (2010), we choose to work with the target variable corresponding to the number of house owners in a given zone for the extensive case. We also select a corresponding intensive variable which is the percentage of house owners, the weights being given by the number of households in the given zone. As potential auxiliary information, we use the covariates presented in Almquist (2010) which are the number (resp: percentage) of non hispanic white, of non hispanic black, of non hispanic asian, of hispanic, of married households with children in the population. The first four percentages are with respect to the population whereas the last one is with respect to the number of households. As far as spatial scale is concerned, we decide to use three different scenarios in the state of Ohio. The first scenario is the disaggregation of the target variable from county level (source) to tract level (target) for the whole of the state of Ohio. Ohio has 88 counties and 2941 tracts so that on average one county contains 33 tracts. The second and third scenario use the county of Franklin as the whole region. Franklin counts 284 tracts, 887 block groups and 22826 blocks. The second scenario is the disaggregation of the target variable from tracts (source) to block groups (targets) in the county of Franklin. In this case, one track contains on average 3 block groups. The third scenario is the disaggregation of the target variable from tracts (source) to blocks (targets) and in that case one tract contains on average 80 blocks. A particular feature of these scenarios is that in all cases the target zones are nested within the source zones.

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

3. METHODS

Let us first briefly summarize the methods we will compare. For more details about these methods, we refer the reader to Do et al. (2014). We focus on the family of simple regression methods which include as a special case the dasymetric methods (see Do et al., 2015).

The value of the target variable Y on a given subzone A is denoted by Y_A and similarly for other variables. For the source zones S_s ; $s = 1,..., S$ as well as for the target zones T_t ; $t = 1,..., T$ the notations S_s and T_t will be used instead to designate a generic source and a generic target. Because of the nested nature of our case here, the intersections between sources and targets coincide with targets.

As shown in Do et al. (2014), the correspondence between intensive and extensive is as follows. It is possible to associate an intensive variable to a given extensive variable by the following scheme. If *Y* is extensive, and if w_A is a weighting scheme of the form

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

where w_k *,k=1,2,...* form a partition of a region Ω, and where *Z* is another count variable, the variable

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

is intensive since

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

Reversely, if one starts from an intensive variable *Y* with weighting scheme w_A of the form (1), it can be transformed into an extensive variable by inverting equation (2). In our case, Y_A is the number of house owners in subregion *A* and $\tilde{Y}_A$ is the percentage of house owners in subregion *A* while Z_A is the number of households of subregion *A*.

Given one auxiliary information *X*, the general dasymetric method predicts the value of Y_t for $t \subset s$ as

$$\hat{Y}_t^D = \frac{X_t}{X_s} Y_s. \tag{3}$$

The formula is the same for intensive and extensive variables but one uses an auxiliary information of the same nature (intensive/extensive) as the target variable. Note that when the auxiliary information is simply the area of the zone, the method is called areal weighting interpolation and it is rather meant for an extensive variable. Note also that areal weighting interpolation applied to an intensive variable is actually equivalent to the dasymetric method applied to the corresponding extensive variable with auxiliary variable being given by its weights.

For regression methods, the type of regression will differ according to the intensive/extensive nature of the variable. For an intensive variable, a gaussian linear regression is used (Flowerdew and Green, 1992) whereas for an extensive variable, one uses rather a Poisson regression (Flowerdew and Green, 1989). As for dasymetric, one uses auxiliary information of the same nature (intensive/extensive) as the target variable.

Let us briefly recall the steps of Poisson regression for an extensive target variable. The distributional assumption is that the mean of Y depends linearly on a set of auxiliary variables X^i , $i=1,2,...,p$ known at target level:

$$Y_A \sim \mathrm{P}\left(\sum_{i=1}^{p} \alpha_i X_A^i\right) \tag{4}$$

A regression of the set of source values based on model (4) is performed at source level yielding estimates $\hat{\alpha}_i$ and used at target level to predict the target values by

$$\hat{Y}_t^{REG} = \sum_{i=1}^{p} \hat{\alpha}_i X_t^i \tag{5}$$

For the intensive target variable case, the distributional assumption, as presented in Flowerdew and Green (1992), is that

$$Y_A \sim \mathrm{N}\left(\sum_{i=1}^{p} \alpha_i X_A^i \, , \sigma^2/n_A\right) \tag{6}$$

where n_A are known and represent the number of underlying points in A. It is thus an intensive variable for which the weights at target level are given by $w_{t:t\subset s} = \dfrac{n_t}{n_s}$. Based on model (6) and on these aggregation weights for Y, one gets the following regression equation at source level

$$(Y_1,...,Y_n)' = WX(\alpha_1,...,\alpha_p)' + \varepsilon \tag{7}$$

where W is the S×T matrix with elements $w_{t:t\subset s}$ and X is the matrix with elements x_{tj} being the values of auxiliary variable j on target t. The estimation of this model yields estimates $\hat{\alpha}_i$ used to predict the target values by

$$(\hat{Y}_{t_1}^{REG},...,\hat{Y}_{t_T}^{REG})' = X(\hat{\alpha}_1,...,\hat{\alpha}_p)' \tag{8}$$

Do et al. (2015) introduce a so called scaled regression by constraining the Poisson regression to enforce the often quoted pycnophylactic property. It is the property of preservation of the initial data in the following sense: the predicted value on source S_s obtained by aggregating the predicted values on intersections with S_s coincides with the observed value on S_s:

$$\sum_{t:t\subset s} \hat{Y}_t = Y_s$$

For the Poisson regression, we get the following correction of the regression predictor, for $t \subset s$:

$$\hat{Y}_t^{SCL} = \frac{\hat{Y}_t^{REG}}{\hat{Y}_s^{REG}} Y_s \ , \tag{9}$$

where $\hat{Y}_s^{REG}$ is simply given by the aggregation rule $\hat{Y}_s^{REG} = \sum_{t \subset s} \hat{Y}_t^{REG}$. We propose here to do the same for gaussian regression of intensive variables by using the empirical conditional expectation of Y_t given the source values, i.e. for $t \subset s$

$$\hat{Y}_t^{SCL} = \hat{\mathrm{E}}(Y_t \mid Y_1, Y_2, \ldots, Y_s) = \hat{Y}_t^{REG} - \hat{Y}_s^{REG} + Y_s \, , \tag{10}$$

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

A very important point for the sequel, which is shown in Do et al. (2015), is that the dasymetric method with a given auxiliary information is equivalent to the scaled Poisson regression with this auxiliary variable as unique regressor without a constant .

Another approach for this problem is to use an EM-algorithm strategy as done in Flowerdew and Green (1991) for the Poisson case and in Flowerdew and Green (1992) for the Gaussian case. Indeed the areal interpolation can be considered as a missing data problem with target values as missing data. We can summarize the steps as follows: the expectation step (E-step) is either (9) or (10) and yield values for the targets and the maximization step is the regression at target level based on models (4) or (6).

4. RESULTS

Table 1 summarizes the notations used for presenting the results. To illustrate the meaning of this table, let us take two examples. The indices have two or three positions: for example "Dhh" or "I.Dhh". When necessary, an additional index in position one will indicate either the intensive/extensive nature or the spatial support depending upon background.

In positions 2 and 3 (potentially after the dot), an index "Dhh" means that we are using the *dasymetric* method (D) with the auxiliary information *percentage of households* (hh). An index "I.Dhh" specifies moreover that it is for the *intensive* target variable (percentage of house owner).

In positions 2 and 3, an index "Ebe" means that we are using the regression method with the *EM* algorithm (E) for the *best* model choice (be) (independent variables are chosen by AIC criteria). An index "B.Ebe" specifies moreover that it is for the with *blocks* as target zones.

Table 1. Notations

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

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

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

4.1. Auxiliary information selection

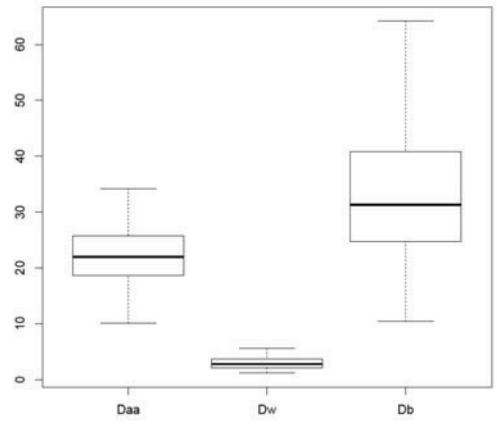
The choice of a good auxiliary information is an important question for areal interpolation in practice. First of all, it is unclear whether a choice of variables which is good for the regression step will be the best for predicting target values. On the other hand, it it difficult to devise a prediction-targeted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

Table 2. Performance of the dasymetric method for each auxiliary information

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

We compare several dasymetric methods obtained by using the different auxiliary information at our disposal using scenario 1 (Ohio) for the extensive case. Table 2 displays the corresponding median error criterions showing that it is very important to select the best auxiliary information since the relative error can vary from around 3 percent to 30 percent. The second row displays the correlations and the non monotonicity of these numbers shows that one should not trust correlation to select an auxiliary information. Figure 1 presents the boxplots of the source errors for the best dasymetric (here: based on the number of whites), the worst dasymetric (here: the number of blacks) and an intermediate case corresponding to areal weighting interpolation. We see that not only the best choice outperforms the other ones by far but also that the variability accross sources is quite high for these choices.

Figure 1. Dasymetric methods for Ohio - extensive approach



4.2. Comparison between different regression methods

In this section, we focus on comparing the different methods using scenario 1 (Ohio) for the extensive case. The implementation of the Poisson regression approach presents some peculiarities. The first one is about the choice of link function. The usual choice for Poisson regression is the logarithm link leading to $\mathrm{E}(Y) = \exp(\sum_{i=1}^{p} \beta_i X_i)$ and it is the so called natural link in this generalized linear model. However we argue in Do et al. (2015) that the identity link is more adapted when relating such extensive variables to auxiliary extensive variables. For example, it seems more natural that the *number of house owners* is proportional to the *population* rather than to be exponentially related to the population. Moreover, empirically, the AIC criterion is 1000 times bigger for the log link specification.

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

Figure 2 presents the boxplots of the counties error criterions for the state of Ohio and for the Poisson regression performed on the *number of house owners*. Table 3 presents the corresponding median error criterions.

Figure 2. Poisson regression methods for Ohio - extensive approach

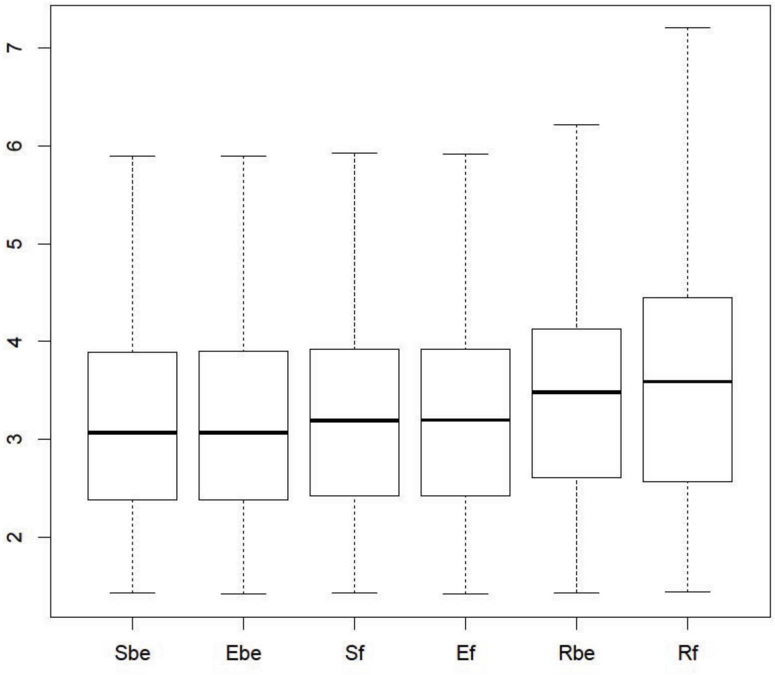


Table 3. Median error criterions - Poisson regressions for Ohio

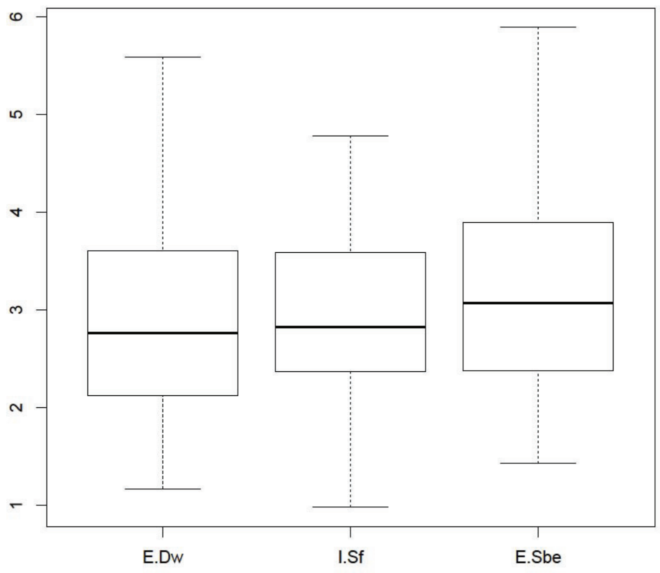
Sbe	Ebe	Sf	Ef	Rbe	Rf
3.069	3.069	3.193	3.196	3.482	3.594

The order of magnitude of the errors is around 3 percent and they are very comparable. The selection of variables strategy selects the model without the variables *married household with children and area* but it seems that keeping a full model does not make a big difference. We see that the scaled regression tends to perform better in general, and even better than the EM approach. However, it turns out that the best of the regression methods gets a 3.069 error criterion and does not outperform the best dasymetric obtained in the previous section with a 2.764 error criterion.

4.3. Intensive versus extensive approach

For the purpose of interpolating the target variable *number of house owners*, we have the choice between two strategies. The first one is to work on the raw variable which is extensive and use a Poisson regression approach. The second one is to work on the *percentage of house owners*, use a gaussian regression approach and transform back the predicted percentages into counts using the knowledge of population on targets if known. In a different situation when this knowledge is not guaranteed, a more complex method is available which disaggregates separately numerator and denominator of this percentage using extensive variables methods. In this section we compare the first two approaches only, the last one giving results very similar to the second one in our case. We use scenarios 1 and 2. For Ohio in scenario 1, Figure 3 shows that the best method is the dasymetric method with auxiliary information given by the *number of whites* applied to the count target variable *number of house owners*. For Franklin in scenario 2, the right panel of Figure 6 shows again that the best result is obtained when working with the count variable rather than the percentage and it is obtained by the regression on the best subset of auxiliary variables. We also see that scaled gaussian regression that we introduced in section 3 is the second best.

Figure 3. Best methods for Ohio - intensive and extensive approaches



4.4. Spatial scale

In this section, we examine the effect of spatial scale on the areal interpolation problem. For this we compare scenarios 2 and 3 on the county of Franklin. Figure 4 (respectively Figure 5) presents the distributions across sources of the error criterion in the case of disaggregation of the extensive variable *number of house owners* (respectively of the intensive variable *percentage of house owners*) at block level and at block group level. Tables 4 and 5 display the corresponding median error criterions.

Figure 4. Methods for the extensive case for Franklin

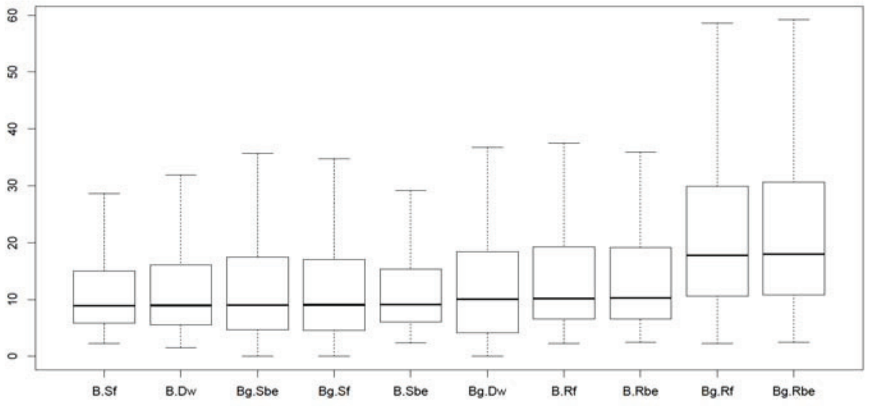


Table 4. Performance of the methods in the extensive case

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

For the extensive case, the best model strategy selects the model without the variables *number of blacks* and *area.* For the intensive case, the best model is the full model.

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

When we compare scenarios 2 and 3 on Figure 6, we see that disaggregation to blocks is more accurate than to blockgroups. Even though the second problem seems easier because the blockgroups are coarser than blocks, one should not forget that the auxiliary information is used at target level resulting in a larger amount of information used for blocks. The medians of source error criterions at block level are thus slightly smaller and the variances are much smaller.

Figure 5. Methods for the intensive case for Franklin

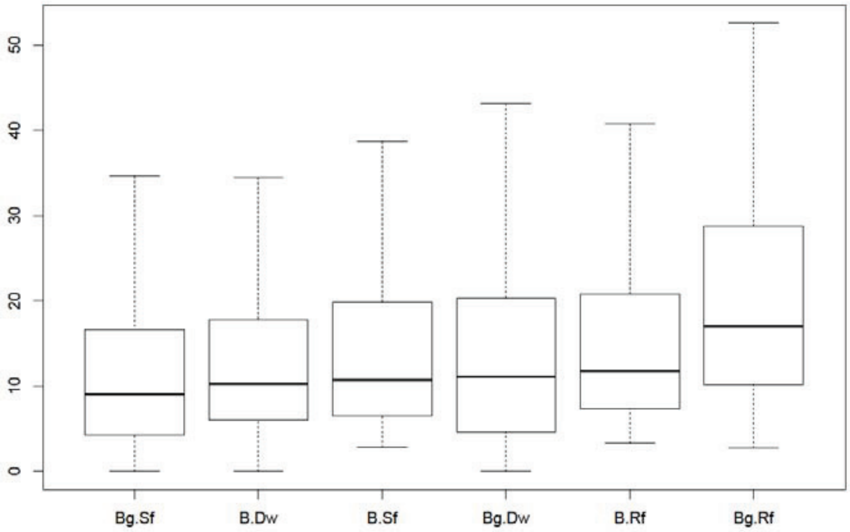
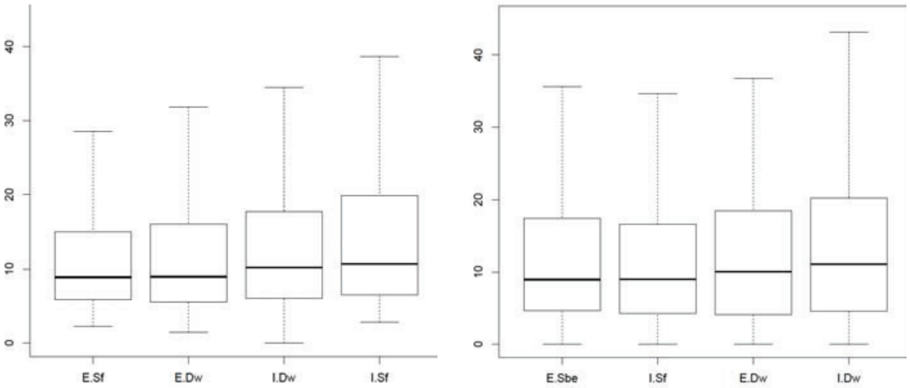


Table 5. Performance of the methods in the intensive case

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

Figure 6. Comparison of block (left) and blockgroup (right) levels for Franklin intensive and extensive variables



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

5. CONCLUSION

We should keep in mind that this study has a particular geometry due to the nesting of targets into sources. In a more general case, some border effects will interplay but we believe that, as long as the size of targets is much smaller than the size of sources (disaggregation), the results should not be very different.

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

REFERENCES

Almquist Z. W. (2010). US Census spatial and demographic data in R: the UScensus2000 suite of packages. *Journal of Statistical Software*, 37, 1-31.

Calcagno V., de Mazancourt C. (2010). glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models. *Journal of Statistical Software*, 34, 12.

Do Van H., Thomas-Agnan C., Vanhems A. (2014). Accuracy of areal interpolation methods: Count data, preprint.

Do Van H., Thomas-Agnan C., Vanhems A. (2015). Spatial reallocation of areal data: a review, *Revue d'Economie Régionale et Urbaine*, Forthcoming.

Flowerdew R., Green M. (1992). Developments in areal interpolation methods and GIS, *The Annals of Regional Science*, 26, 67-78.

Flowerdew R., Green M., Kehris E. (1991). Using areal interpolation methods in geographic information systems. *Papers in Regional Science*, Vol. 70, Issue 3, 303-315

Hastie T., Tibshirani R., Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. *Springer Series in Statistics*. 2nd ed.

Ripley B., Venables B., Bates D.M., Hornik K., Gebhardt A., Firth D. http://cran.r-project.org/web/packages/MASS/MASS.pdf

COMPARAISON DES MÉTHODES DE RÉAFFECTATION SPATIALE DE DONNÉES SURFACIQUES SUR LA BASE DES DONNÉES DU RECENSEMENT DE POPULATION DE 2010 AUX ETATS-UNIS

Résumé - *Le problème d'interpolation spatiale de données surfaciques est celui de la réaffectation d'une caractéristique sur une partition cible de l'espace à partir de la connaissance de cette même caractéristique sur une autre partition appelée source en utilisant de l'information auxiliaire. L'objectif de ce travail est d'utiliser une base de données démographiques 'US census 2010' disponible dans le* package *de R (Almquist, 2010) dans le but de tester plusieurs méthodes basées sur la régression dans le cas de données liées à des comptages. Le fait que les données de cette base soient disponibles à différentes échelles spatiales fait l'originalité et l'intérêt de cette comparaison. Un autre apport original de notre démarche est la comparaison entre l'approche extensive et l'approche intensive pour une variable qui est un rapport de deux variables de comptage. Nous incluons également dans la comparaison la version normalisée de la régression dans le cas extensif introduit dans Do et al. (2015) et avec la version normalisée de la régression dans le cas intensif proposée ici. Finalement nous formulons quelques conseils pour choisir des variables auxiliaires.*

Mots-clés - INTERPOLATION DE DONNÉES SURFACIQUES, PROPRIÉTÉ PYCNOPHYLACTIQUE, DÉSAGRÉGATION SPATIALE