Accuracy of areal interpolation methods 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*

Résumé. L'analyse de données socio-économiques nécessite souvent de combiner des bases de données provenant de différentes sources administratives, données collectées sur plusieurs partitions différentes de la zone d'intérêt. Il est donc nécessaire de transformer les données provenant d'unités spatiales d'origine ("sources") en données associées aux unités spatiales "cibles". Par exemple, on peut s'intéresser à un quadrillage commun régulier d'une zone d'intérêt et souhaiter adapter toutes les informations initiales à cette nouvelle partition cible unique. Cette option est actuellement à l'étude en France à l'INSEE et en Europe avec la directive de l'UE 'INSPIRE' (INfrastructure for SPatial InfoRmation). Ces techniques de transformation de données se répartissent principalement en trois catégories: la méthode des poids proportionnels, les techniques de lissage et l'interpolation par méthode de régression. Nous proposons un modèle basé sur les processus ponctuels de Poisson afin d'estimer l'erreur de prévision de ces différentes méthodes dans le contexte de données de comptage et pour des unités cibles définies par un quadrillage régulier. Nous montrons que l'erreur dépend de la nature de la variable d'intérêt et sa corrélation avec la variable auxiliaire. Nos résultats portent principalement sur la méthode des poids proportionnels et les méthodes basées sur la régression de Poisson et nous montrons qu'il n'existe pas de méthode qui domine toujours.

Mots-clés. Interpolation de surfaces, désagrégation spatiale, Propriété pycnophylactique, Incongruityé spatiale, erreur de prévision

Abstract. The analysis of socio-economic data often implies the combination of data bases originating from different administrative sources so that data have been collected on several different partitions of the zone of interest into administrative units. It is therefore necessary to allocate the data from the source spatial units to the target spatial units. A particular case of that problem is when one uses a common regular grid and re-allocate everything to this single target partition. This option is currently under study in France at INSEE and in Europe with the EU directive 'INSPIRE', or INfrastructure for SPatial InfoRmation. There are three main types of such techniques: proportional weighting schemes, smoothing techniques and regression based interpolation. We propose a model based on Poisson point patterns to study the 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. We focus on proportional weighting schemes and Poisson regression based methods. There is no technique which always dominates.

Keywords. areal interpolation, spatial disaggregation, pycnophylactic property, spatial incongruity, accuracy.

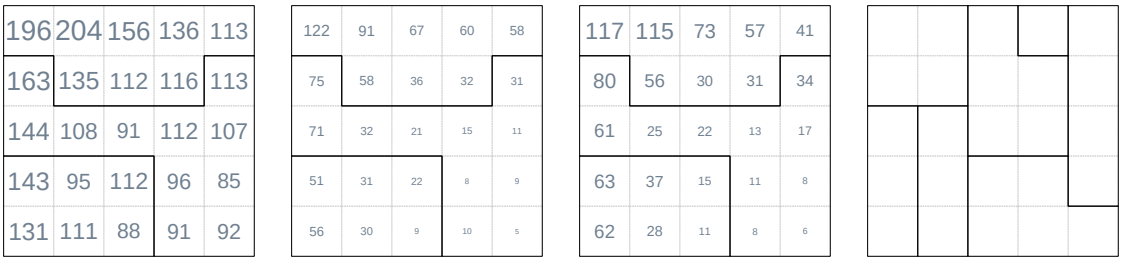
1 Introduction

The analysis of a socio-economic data often involves the integration of various spatial data sources. Those data are often independently collected by variety of offices in use for different purposes. The zonal set systems used by distinct offices are rarely compatible and it leads to many difficulties in the analysis of data. The problem of merging data bases on different spatial supports is called the areal interpolation or basis change (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 methods are proposed to handle this problem. It is difficult to compare the accuracy of the different methods which depends on several factors: nature of the target variable, correlation between the target and auxiliary variables, relative size between two zonal sets,... In this document, we consider the accuracy for count data, which are frequent in the literature, in the case when target zones are nested inside source zones.

2 Count data

We consider the following two types of target variables. In the first case, for a sub-region A of the region of interest, the target variable is a count of points included in A. An example of such variable is the population of a sub-region. In the second case, the target variable is the number of points per areal unit. In the first case, the variable belongs to the family of extensive variables and in the second case to the family of intensive variables (see Do et al. 2013 for more details). In the sequel, we will mainly focus on the extensive case. As we will see in the next section, some methods make use of an auxiliary information, and we imagine that the auxiliary information is also of a count process nature. We introduce a model for this type of variable by assuming that there exists an underlying unobserved Poisson point pattern Z (in the 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 in A. In order to illustrate and evaluate the methods, we will use a simulated toy example. On a square grid with 25 cells, we design three sources and seven targets as unions of cells. On the figure, we see the cell counts for two target

variables Y_1 and Y_2 and for one auxiliary information variable X. The figure also shows the boundaries of targets and sources. We simulate the underlying point process with an inhomogeneous intensity. We then recover the counts at the cell level.



(a) Y_1 at cells (b) X at cells (c) Y_2 at cells (d) Targets

3 Methods

Many methods are used to solve the areal interpolation problem. Do et al.(2013) classify these methods into three groups: smoothing, dasymetric and regression based methods. Goodchild and Lam (1980) present areal weighting interpolation which does not use any additional information: the data is allocated to the targets based on the assumption that the target variable is homogeneous at source level. An alternative class of methods named dasymetric uses an available auxiliary variable instead of area. The value at target level is assigned a fraction of its source value proportional to the auxiliary variable which is supposed to be known at the target level. Voss et al. (1999) use road segment length and the number of nods of roads to calculate the weights of dasymetric methods. Other approaches using auxiliary information are regression based methods. Flowerdew and Green (1992) use information at target level to set up an ordinary linear regression of the source values for house price which is an intensive variable. Flowerdew et al. (1991) use Poisson regression to predict population (which is an extensive variable) with categorical auxiliary information. Goodchild et al.(1993), Yuan et al. (1997) use a two steps procedure with a regression of source values on expert based control zones densities as the first step. The purpose of our paper is 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.

4 Poisson point pattern model

Using the model based on Poisson point patterns we introduced in Section 2, we study the accuracy of these techniques for the case of an extensive count target variable. For simplification reasons, we consider a regular grid of targets T that are nested in sources S. As mentioned before, we assume that there is an underlying unobserved Poisson point pattern Z and that the target variable Y on a subzone A is the number of points of Z in A. For any zone A, Y_A is Poisson distributed with intensity $\lambda_A = \int_A \lambda(x)dx$ and is determined by

$$Y_A = \sum_i \mathbf{1}_A(Z_i)$$

With this model Y_A and Y_B are automatically independent for all disjoint couples of regions A and B. Because the targets are nested within sources, any information at the target level is also an information at the intersection level between targets and sources and this justifies the comparison between dasymetric and regression methods in this framework.

We assume that an explanatory variable X is known at the target level. For this information to be useful, there must be a relationship between the auxiliary variable and the target variable. More specifically, we propose the following relationship

$$Y_T \sim \mathcal{P}(\alpha|T| + \beta x_T)$$

and estimate the parameters α and β using Poisson regression. We restrict attention to univariate regression because dasymetric method only use univariate auxiliary information.

5 Accuracy: methodological assessment

5.1 Areal weighting and dasymetric

Some authors study the accuracy of areal interpolation methods. Most of them follow an empirical approach using a benchmark data such as Gregory (2002). A few authors such as Sadahiro (2000) are concerned about a methodological approach: he considers the accuracy of the weighting interpolation and point-in-polygon method based on a stochastic distribution of points.

In our paper, we assess the accuracy by considering mean square errors for each method under our model. We show that the error depends on the nature of the target variable, correlation between the target variable and the auxiliary variable and comparative sizes between sources and targets. In particular, we see that the mean process is driven by two effects: the $\alpha|T|$ term of the mean reflects the effect of the area and the βx_T reflects the effect of auxiliary information. For the two approaches of proportional weighting

schemes, we prove that dasymetric dominates when the effect of auxiliary information is stronger, and areal weighting interpolation dominates when the target variable is quite homogeneous.

We also propose a combination approach that takes care of both effects of the auxiliary variable and homogeneity. This method reduces the error in comparison not only with the two former proportional methods but is also optimal in the family of linear predictors. However, this predictor cannot be implemented without the knowledge of the parameters of the model but it is an interesting tool to understand the relationship between proportional methods and regressions.

5.2 Poisson regression and dasymetric

We compare asymptotically the Poisson regression predictor $\hat{Y}_T = \hat{\alpha}|T| + \hat{\beta}x_T$ and the composite predictor where the unknown parameters are replaced with an estimation obtained by Poisson regression (Poisson composite predictor hereafter). We show that the Poisson regression method asymptotically satisfies the total pycnophylactic property (preservation of the total count at the region level) and yields an asymptotically unbiased predictor. On the other hand, the Poisson composite predictor satisfies the classical pycnophylactic property (preservation of the count at each source level). We show that the difference between both predictors converges to zero in probability.

6 Accuracy: simulation assessment with a toy example

We consider two cases of relationships between the target variables and auxiliary information variables. For the same auxiliary information variable, we simulate two different target variables: in one case, the effect of the auxiliary variable on the expected value of Y is dominant and in the other case, the dominant effect is that of the area. The results confirm our theoretical findings about dasymetric and areal weighting. We also compare on this example Poisson regression with areal weighting and dasymetric method.

Bibliographie

[1] Do Van Huyen, Christine Thomas-Agnan, Anne Van Hems (2013) Spatial reallocation of areal data: a review, summited to 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] 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.
[8] Sadahiro (2000), Accuracy of count data estimated by the point-in-polygon method, *Geographical Analysis*, 32(1).
[9] 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.
[10] 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.