Transferring Indicators into Different Partitions of Geographic Space

Christine Plumejeaud[1], Julie Prud'homme[2], Paule-Annick Davoine[1], Jérôme Gensel[1]

[1] LIG, Laboratoire d'Informatique de Grenoble
681 rue de la Passerelle, Domaine Universitaire BP 72,
38 402 St Martin d'Hères cedex, France
{Christine.Plumejeaud, Jerome.Gensel, Paule-Annick.Davoine}@imag.fr
[2] UMR 6012 ESPACE CNRS – Université d'Avignon,
74, rue Louis Pasteur, 84029 Avignon cedex 1, France
jul.prudhomme@gmail.com

Abstract. Nowadays, spatial analysis led on complex phenomenon implies the usage of data available on heterogeneous territorial meshes, that is to say misaligned meshes. Then, combine these data requires the transfer of each dataset into a common spatial support that can be exploited. This is known as the Change Of Support Problem (COSP). However, it appears that transfer methods are numerous, and they are often linked with a regression model, and other parameters whose selection and tuning may not be straight forward for a non-expert user. Furthermore, the process is also very dependent from both the nature of the data to be transferred and their quality. This paper first proposes a brief overview of some available transfer methods, giving the premises for the characterization of each method. A use case illustrates a transfer operation, and reveals its main difficulties.

Keywords: Territorial mesh – spatial analysis – statistical datasets – downscaling – aggregation – interpolation.

1 Introduction

Very often, spatial geo-referenced data related to a certain phenomenon of interest (sociological, environmental, or political one) are collected on a support (that is to say a set of geo-referenced locations) which is not the one on which the phenomenon is actually active, observable, or even studied. For instance, temperatures are measured at some punctual locations, whereas this phenomenon is a continuous field, operating on the ground of the Earth, modeled here as a planar surface. Moreover, considering the tremendous increase of the number of data sources of these last past years, researchers in Social Sciences or Geography often have to combine data coming from different kinds of support. For instance, in order to measure the green impact of human's activities, one need to combine data associated with economic basins together with gridded land cover data extracted from some processing on spatial images. In those cases, the analysis and the interpretation of data require the harmonization of data supports, that is to say to transfer data inside a common support so that heterogeneous data and supports can be compared. These problems are identified in the class of *Change Of Support Problems* (COSP) [1].

A lot of previous works have studied and proposed methods for transferring data between different supports [16], [17], [24]. Usually, those studies focus on the side effects of the proposed transformation. Those side effects are mainly caused by the change of scale and the change of shape of the support. This problem is known as the MAUP (*Modifiable Areal Unit Problem*) [25], [9]. However, to our knowledge, there is no system offering a classification and a characterization of those methods, which could handle a part of the advanced knowledge in spatial analysis or in mathematical fields of expert users. Yet, some attempts into this direction have been made [16], [0], [10].

Our global objective is to show that such classification and characterization of transfer methods can be exhibited, in order to automate their execution by an interactive system dedicated to expert users. This paper aims at exposing our ideas and a methodology allowing the automatic transfer of data between heterogeneous supports.

First, we make a brief survey of spatial analysis methods that can be used to transfer data from one support to another. Then, we identify a list of basic characteristics that should be known and selected by the user (cost calculation, complexity of the algorithm, side-effects and parameters of the method, etc.), together with their meaning. Finally, the estimation of population performed on the districts of the city of Grenoble in France, illustrates the transfer mechanism through the examples of two methods.

2 The Change Of Support Problem

Geographical information is associated with a set of locations that can be modeled using objects, characterized by their geometry and some coordinates described into a mathematical space. For this space, a metric is supplied in order to get a measure of the interaction between objects. This set of mathematical objects is known as the *support* of information. Generally, and for sake of simplicity, the mathematical space that is used is a planar Euclidian space, that is to say a Cartesian 2-space, the Euclidian distance function being then used as length-metric. Nevertheless, many interesting methods, based for example on spherical spaces [29], and/or other kind of distances, can express the interaction between locations in a more suitable way when human geography is concerned. For instance, the use of a road distance-time matrix between locations, or of an orthodromic distance, allows a more realistic analysis. But most of the methods that are defined on specific spaces or specific distances require intensive computing, and have not been disseminated yet inside the GIS community [23].

In a planar Euclidian space, one can consider various kinds of support: points, lines, or surface. In our study, we limit our proposal to data linked to surfaces (often called "areal zonings"). In fact, a *tessellation* is applied to the projection of the geographic space onto a plane surface. This tessellation constitutes the support, that is to say a gridded representation of a plane surface into disjoint polygons. These polygons are usually squared (raster), triangular (TIN), or hexagonal ones. A 2D tessellation involves the subdivision of a 2-dimensional plane into polygonal tiles (polyhedral blocks) that completely cover the space. The term *lattice* is sometimes used to describe the complete division of the plane into regular or irregular disjoint polygons. A lattice is said to be regular when one can observe the repetition of some identical shape over the space. For example, a *grid* is a regular polygonal lattice. On the opposite, a *mesh* is a kind of irregular lattice, composed of polygons that have various shapes and size. We use the term of "*discontinuous mesh*" in order to designate an irregular tessellation of space, that does not fully cover the plane. This case occurs, for instance, when defining some employment basins on a given territory: basins limits usually form a lace of this territory.

Some meshes may be nested into other meshes: then the upper level mesh can be built from territorial units of the lower level, using aggregation rules that have to be established. For instance, the Nomenclature of European Territorial Units (NUTS) relies on a hierarchical structure: units of NUTS2 level, corresponding to regions (*régions* in France), are built by aggregation of units of NUTS3 level, corresponding to departments (*départements* in France). If the hierarchy is *strict* (each unit has only one superior unit) and *onto* (any unit of a non-elementary level is composed of at least one unit of the lower level), the transfer of absolute quantitative data from lower levels to upper levels is made straightforwardly by additive operations (addition, maximum, minimum) on data accounted on each unit of the lower level [31]. Finally, two meshes are said to be *misaligned* if one is *not* nested in the other and if their boarders cross each other. This is the case when considering, for instance, administrative units and river catchments on a given territory (a portion of space). In general, they are misaligned because, at some level, a river crosses several administrative units, and its catchments may overlap with several administrative units. In this study, we focus on methods for transferring quantitative data (whether they are absolute or ratio) between two misaligned meshes. Indicator and variable are here considered as synonyms and refer to a property which is considered of interest for any territorial unit of the support.

Based on the vocabulary introduced by [21], the term "change of support" is used when some data is known on a source support, and this data is transferred onto a target support, using some methods whose complexity is variable. This problem is also known as « *spatial rescaling* » [24], or « *downscaling* », since the major difficulty lies in the disaggregation of a variable on a finer scale support that is misaligned with the source support. Transfer techniques are also named « *cross-area aggregation* » [14], or « *areal interpolation* » [11], [12], or « *polygon overlay* ».

The relative scale of meshes is a substantial aspect of the transfer: from one mesh to another, the average area of each unit (or cell) differs, and even inside a given mesh, the degree of variability of these areas is not constant also. When the average area of units of a mesh A is lower than the average area of units of a mesh B, A is said to be finer that B. When most of the units of the mesh A are nested inside units of the mesh B, the situation allows us for aggregating A values into B, with a margin of error, due to the little sub-set of A units that may cross the boarder of B.

Transfer methods from one support to another are thus based on the establishment of an intermediate support, nested inside the target mesh, in the case where this mesh is finer than the source mesh. Then, possible operations are:

- *Disaggregation* which reallocates data on objects of smaller size (either punctual or polygonal), fully included into the units of the source support and into the units of the target support, through the creation of a least common spatial denominator between the source and the target supports.
- *Aggregation* which cumulates data of the source units into the target units, following a geometric inclusion rule. This operation is used after a disaggregation in order to get an estimation of the variable on a misaligned mesh, or when passing from a lower scale to an upper scale in case of hierarchical organization.
- *Interpolation* which builds a density surface using data associated with a punctual or a polygonal support. The support of this density surface is a regular mesh as finer as possible.

Interpolation is generally integrated into a sequence of operations, where data is first reduced to a seedling of points, then interpolated, and finally re-aggregated into the target support. Indeed, the amount of data accounted on a polygon can be considered as concentrated into the center of this polygon, which can be the centroid, or a point inside the corresponding unit that have been selected by the user due to its fitness regarding the study needs: the biggest town, the lowest altitude point, etc. In general, the computed surface aims at giving a continuous representation of discrete or punctual measures. Since this surface is associated with a fine discretization of the space, on which one can sum data of cells, according to their spatial inclusion into the target units, this enlarges considerably the set of available methods for transferring data between misaligned supports.

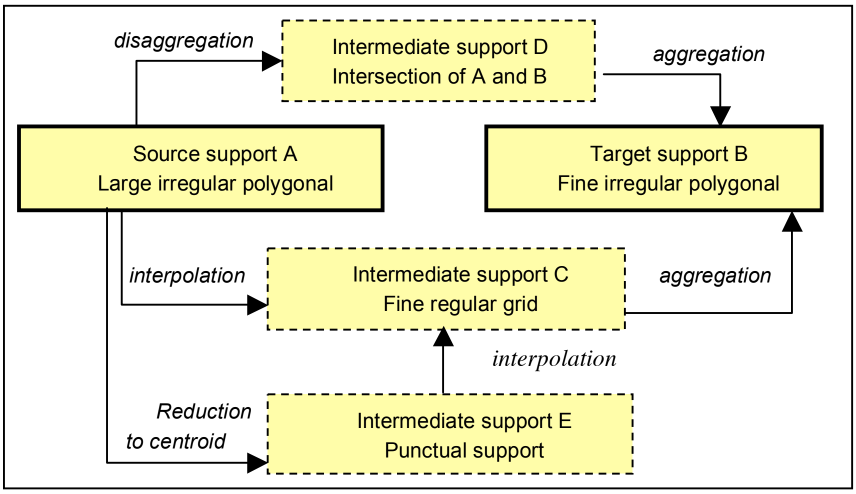


Fig. 1. Support transformations for data transfer.

Figure 1 shows that these three operations that are closely linked to the nature of both source and target supports. All these transformations have also a side effect on the analysis of data, due to the change of size and shape of the support. This is mainly due to the lost of variance on data when grouping units into larger units, which is known as the aggregation or scale effect. Also, alternative formations of units produce different results and inferences: this is called the grouping or zoning effect [25].

We present in the following subsections a synthetic overview of various transfer technique families, and we try to link them with a set of useful criteria for the user. The characterization of the support, for both source and target, can be helpful in the selection of one particular kind of transformation. Yet, the type of variable to be transferred also plays a key role: a relative variable (a ratio) cannot be manipulated in the same way as an absolute quantitative variable (that is to say a count on units), in particular through aggregation operations. The sum of lower units rates is not equal to the upper unit rate. There are few methods adapted to ratio manipulation, and often, one has to operate separately on the both components of a ratio. In the same way, some ancillary data can be used to adjust the transfer, considering that those data may have a spatial correlation with the data to be transferred.

2.1 Aggregation

About aggregation methods, we could have mentioned the regionalization method, the optimization and simulated annealing methods [26], or the clustering methods. However, these techniques aim at building a new mesh according to constraints given by the user (similarity, homogeneity, continuity, etc.) But, since our goal is to transfer a variable towards a target mesh whose borders are defined, we focus on the technique of data-matrix aggregation. Data-matrix is data supported by a grid. In Figure 1, it corresponds to the transformation of the support C into the support B. The idea is to aggregate variables, which are known on a grid support into a target mesh. By superimposing a target mesh on the grid support, one can accumulate the values of the cells considering their spatial inclusion into units of the target mesh, and thus determine the value of the variable in the target support, within a margin of error. This error depends on the weight of the cells of the intermediate mesh and on the process used to reallocate a cell to a unit of the target-mesh. For instance, if the cell mostly belongs to the polygon of the target mesh, the total weight of the cell is assigned to the cell itself, or else it is reallocated proportionally to the surface of the intersection between the cell and the polygon. To reduce as much as possible the error, the intermediate mesh must be as fine as possible. Also, much of the variability in the information is lost since information is re-aggregate in bigger sized and differently shaped units (scale and zoning effect). This kind of method is widely used in the environmental domain – where data on raster support abound – when data must be transferred into socio-economic meshes like the NUTS (Nomenclature of Territorial Units for Statistics) [15], [30].

2.2 Disaggregation

Simple Areal Weighting.
This method enables the transfer of variables between two non-nested meshes by means of an intermediate support D, resulting of the intersection between the source and the target supports. It assumes the uniformity of the density of the variable in each unit. The ventilation of the variable is done proportionally to the surface of the intersection between the source mesh and the target mesh. It is easy to implement and does not require a big data sample. The algorithm is available in most of the GIS. However, the strong assumption of the uniformity of the density of the variable prevents the method from taking into account the local variations of data inside units.

Modified Areal Weighting – Regression.
This method assumes that a spatial correlation holds between a set of variables [7]: one must identify an ancillary data (called *predictor*) spatially correlated with the variable to be transferred, and whose distribution must be known on both the source and the target supports. Then, the reallocation of the variable on the intermediate support D integrates an assumption about the similarity of the distribution (spatial correlation) between the variable and the ancillary variable (the predictor). In a first step, the spatial correlation model has to be established on the source support. The second step solves a linear regression between the predictor (whose distribution is known on the target support) and the variable to be estimated. An example of this method is provided by [13]. This study highlights that the choice of the regression form between the two variables can be crucial for results accuracy.

Modified Areal Weighting – Control Zones.
This is an enhanced version of simple areal weighting, which make use of an ancillary data known on control zones, which constitutes a third support in addition of the source and target support [16]. This third support can be a set of geographic objects (like buildings for example) that forms a finer mesh whose units are almost all included into both the source and the target support units. Generally, density of the ancillary data on those control zones is said to be constant [19], and through a mathematical expression of the relation existing between the ancillary data and the data to be transferred, it is possible to compute the data on those control zones, and then to re-aggregate into the target support. This method is frequently used because its results are considered to be close to reality. For example, in environmental domain, the trend is to use land use data issued from regular grids composed of pixels of 1ha as control zones [15], [30]. Indeed, numerous variables are directly linked to the land use category: population lives on urban areas, pesticides are spread on agricultural areas, etc.

2.3 Interpolation

The purpose of interpolation is to give the value of a function *F(x,y)* for any point *M* whose coordinates are *(x,y)*, using a sub-set of observed values f_i at locations M_i, where *i* varies from *1* to *n*. Interpolation builds a continuous surface, based on a fine discretization of space. Depending on the algorithm used, discretization can result in a regular grid, or a triangularization of space. This allows then to aggregate computed values into the target units, by summing values associated to cell grids or triangles, or whatever polygonal shape, that are included into to each target unit. There are many methods [2], and [35] lists the following criteria to distinguish between them:

- The property of the algorithm: *determinist* or *stochastic*. In the first case, the estimation tries to adjust a mathematical function with regard to a set of samples, whereas in the second case, a probabilistic rule models the phenomenon and provides directly an estimation error.
- The accuracy of the estimation: sometimes these algorithms build a mean density surface, which does not pass through all samples f_i. The ones that compute a value $F(x_i, y_i)$ equal to f_i are said to be *exact*.
- The spatial extent: this makes the difference between *global* or *local* methods. In the first case, a function models the phenomenon on the whole area, whereas in the second case, estimations are computed within the local neighborhood of each sample f_i.

The current presentation is non exhaustive, and we describe here just a sub-set of available methods, in order to illustrate the criteria listed above. Some kinds of methods are just mentioned, with their references. Also, we do not describe some descriptive methods aiming at discovering spatial or spatio-temporal correlation between variables, using either global indices (like Moran), or local indices (like LISA). Indeed, those methods do not make any estimation, even though they are very useful to find spatio-temporal cluster of similar data, and can help also in the determination of ancillary data.

First, we present *parametric regression* methods since many other spatial analysis methods end the steps of their calculus by a linear or a polynomial regression. The produced surface will be described by a unique formula (they are global, deterministic and exact methods). In fact, in those cases, the surface or curve ignores micro-fluctuations, and generalizes the phenomenon. However, the algorithm is simple, and many libraries propose scripts or methods to compute it.

Trend Surface Analysis (Global Polynomial Surface).
Based on a linear combination of polynomial functions, whose degree must be chosen by the user, this interpolation method corresponds to a least square approximation. A linear regression is a special case where the degree is equal to one. Thus, the surface is modeled by:

$$Z = F(x,y)=b_0+b_1x+b_2y+b_3xy+b_4x^2+b_5y^2+\ldots$$

Parameters of the model (b_0, b_1, b_2, b_3, b_4, b_5, …) are computed by a least square method, which aims at minimizing the error between the predicted value *Z* and the samples f_i by minimizing the sum (1) :

$$\sum_{i=1}^{n}(Z_i - f_i)^2 \ . \tag{1}$$

In Figure 1, this method follows the step A to E (centroid reduction), and then E to C (surface computing). The choice of the degree determines the cost of the computation, both increase with the degree, and the interpretation complexity. Most often, the chosen degree varies from 3 to 5. Like any parametric regression, this method is not well suited in cases where data distribution is sparse and irregular through space [35].

Non-parametric regression methods involve nearly no a priori assumptions about the underlying functional form. At each point M of the support C, such models compute a local distribution function *$F_i(x, y)$*, based on a weighting of f_i by a coefficient λ_i, assuming that such weights decrease with the distance to the point M (2).

$$F_i(x,y) = \sum_{i=1}^{n}\lambda_i f_i \ . \tag{2}$$

This hypothesis is based on the first law of geography stated by Tobler: "Everything is related to everything else, but near things are more related than distant things". The following constraints remain valid for all methods: the sum of weights is equal to one, each weight is always positive or null, and their value depends on the distance between M and the sample f_i. Such models are much more flexible, due to their local nature, and are capable of capturing subtle spatial variations. However, until recently non-parametric models have not been as widely applied as parametric models because of their high computational cost and their great statistical complexity. But thanks to the advances of intensive computing, research is very active in this domain. In this category, one finds:

- Splines [8], which are based on a polynomial regression by pieces, with a constraint of continuity, derivability and smoothing of the surface.
- Kernel smoothing methods - (for which many implementations and variations have been published, such as Potential Maps [27]) – which assume that a shape (Gaussian, exponential, disk, etc.) controls the distribution around a neighborhood of each computed point M, decreasing with distance. They require choosing a scope, defining the mean average distance of the influence of each point on its neighborhood.
- Regression with Topological maps, that are a kind of neural-network based algorithms (Self-Organizing Maps) [3].

We present here a basic one, the Inverse Distance Weighting, and some variants that are very often implemented inside GIS, providing a quick overview of spatial variability, identifying zones with high or low values.

Inverse Distance Weighting (IDW).

The IDW method belongs to the family of spatial filters (or focal functions), being determinist and local. Like kernel smoothing methods, a shape for the phenomenon diffusion is selected by the user, and decreasing with distance. IDW assigns to each λ_i of (2) a value that is proportional to the inverse of the distance d powered by n, d being the distance from the computed M to the sample f_i, as shown in the equation (3)

$$\lambda_i = \frac{k}{d_i^n} \ . \qquad (3)$$

Triangular Irregular Network (TIN) is a particular case of *IDW*, where $n=1$ in (3), and every λ_i is null excepted for the three nearest neighbors: those neighbors are defined by a triangulation of space based on *Thiessen* polygons.

For *IDW* computing, a very fine grid is superimposed onto the source support, (see step A to C on Figure 1), and each cell of C is initially valued with a value proportional to the intersected area of units of the support A. Then, $F_i(x,y)$ is computed in each cell, taking into account the neighborhood through a given scope. This is a very fast computing method, but producing maps with a "Bull's-eye" effect around the point of measure. The Shepard's method [33] and its derivates adapt this method, in order to suppress the "Bull's-eye" effect by using a local least square regression.

Pycnophylactic Interpolation.

The pycnophylactic method is another deterministic method, local but exact. Proposed by [34], it aims at producing a smoothed surface from data known on areal zoning, but preserving the total mass of data accounted on source units. It looks like spatial filtering into its first step: a very fine grid is superimposed onto the source support, (see the step A to C of Figure 1), and each cell is initially valued with a value proportional to the intersected area of units of the support A. Then a filter window is moved along the grid to smooth the values, computing the average values of adjacent cells (4 or 8 cells). However, at each iteration, this method constrains the new accounts in order to guaranty the continuity of the surface, as well to preserve the mass accounted on source units. The process ends when a limit for precision is reached (defined by the user). The main drawback of this method is to rub out sudden changes (e.g. local pockets of high population density within a cell) [28], because it works by assuming a uniform variation of the variable density.

In addition, all deterministic methods have two major defaults: first, it is impossible to get an idea on the estimation reliability; secondly, the knowledge about the spatial distribution of measures is not used. *Geostatistical methods* (amongst which Kriging, and all its variants) provide a model of the error, and take benefit of the knowledge about spatial distribution of samples.

Kriging.

Kriging is a global and exact stochastic method. Very frequently used and implemented in GIS, it avoids the MAUP. A South-African engineer, M. Krige, has invented this technique, in order to compute the spatial distribution of gold from a set of drilling mining (punctual support). Matheron [22] has formalized the approach with a mathematical theory: it computes the best linear unbiased estimator, based on a stochastic model of the spatial dependence, quantified either by the variogram or by the expectation and the covariance function of the random field. The first step establishes a model of the spatial distribution, based on the variance and a distance function. The second step integrates this spatial similarity model into a simple linear regression model. There are various kriging forms, depending upon statistical characteristics of studied variable: stability through time, heteroscedasticity, etc. [6]. The main drawback of this method is that it requires numerous samples as input. Furthermore, many datasets do not have a clear spatial dependence model.

2.4 Synthesis

Some common characteristics or parameters should be outlined amongst the various transfer methods. First, considering simple aggregation or disaggregation methods, the strategy of reallocation of the variable to the units of the target support is very dependant on the nature of the variable: for a continuous variable, such as forest land use, one should use the areal weighed reallocation. But for a policy related variable, such as number of hospitals, a reallocation based on the major belonging of units sounds more appropriate.

Interpolation methods, and even complex disaggregation methods, often use regression model based on a function which models the diffusion phenomena, that the user has to provide, based on its own assumptions and knowledge. This is the case for kernel smoothing methods, disaggregations by regression, or focal functions. The user has to provide a scope (the mean neighborhood to be accounted for estimation) for those methods. Then, the user has to select a type of distance (simple contiguity, Euclidian, orthodromic, etc.) and be aware of the topology of the space.

The user may like to express some preferences. For example, only geostatistic methods (such as kriging) provide automatically an error estimation. For the others, the user should compare results of various models to each other in order to check whether convergence can be reached, which is a more tedious task. The user may also give the priority to the computing time: global regression methods or focal functions are generally very efficient, whereas kernel smoothing methods are much more time consuming.

3 A Case Study: the City of Grenoble, France

3.1. Context

In the city Grenoble, the partition (exactly the tessellation) that is used for urban development, political action and spatial planning is the district mesh. Composed of 20 districts, this district mesh corresponds to the actual structure of the urban area of Grenoble, since each unit presents of certain degree of homogeneity, in terms of buildings, urban activities, and the socio-demographic profile of population. But this mesh, created during the eighties, is misaligned with the IRIS mesh, which is a prior official mesh used by the INSEE (the French national institute of statistics) to publish statistical accounts of population (see Figure 2).

Knowledge about population of districts is a crucial problem when dealing with planning urban development. Then, we have to cope here with an estimation problem, that requires the selection of the best-fitted method. Here, two question are raised: amongst all methods previously presented, which ones could be applied? Are their results equivalent?

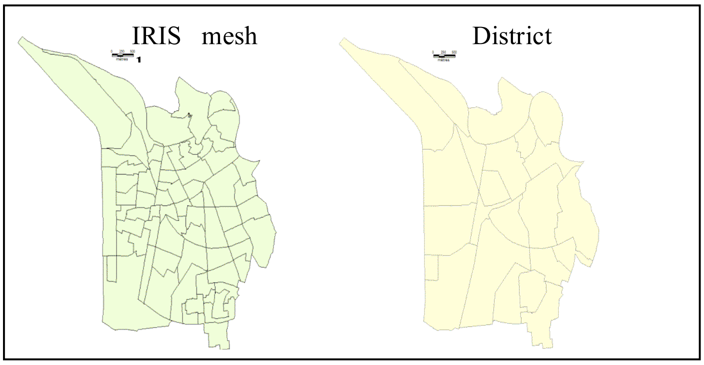


Fig. 2. Different partitions of the city of Grenoble.

Clearly, in this case study, data supports are misaligned meshes (IRIS being considered as the source support, and district mesh as the target support). In order to achieve the transfer of data from IRIS to the district mesh, the technical municipality services in charge of the urban management and planning make use of a simple areal weighting method. The reallocation rule assigns accounts of one source unit to the target unit which overlaps it the most. Accounts on districts are obtained by the summation of IRIS units accounts that are, in the majority, spatially included into target units. As an example, the *Capuche* district (a central district on Figure 3) illustrates this algorithm. Composed of many IRIS units, which are not always fully included inside the target unit, the estimation of the population of the *Capuche* district comprises a great number of persons living in buildings belonging to another district. This increases in a tremendous way the estimated account of population on the *Capuche* district.

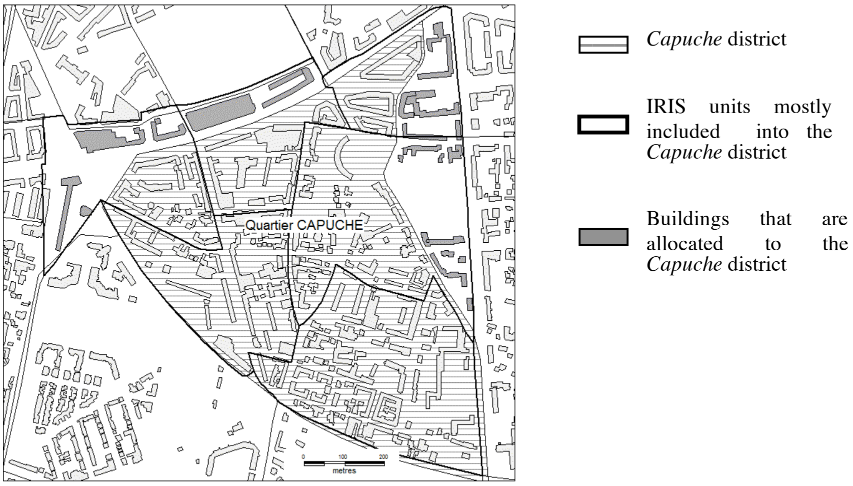
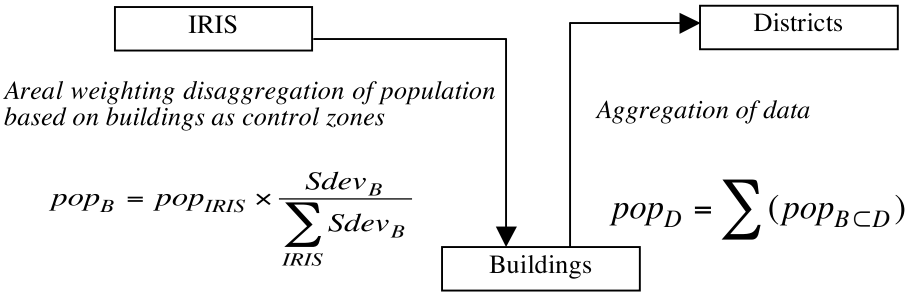

Quartier CAPUCHE
0 100 200
metres
Capuche district
IRIS units mostly included into the *Capuche* district
Buildings that are allocated to the *Capuche* district

Fig. 3. Overestimation of population accounts using the simple areal weighting on the *Capuche* district.

This example shows the limits of a transfer method based on the simple areal weighting disaggregation. It allocates very frequently some source units to one single target unit, whereas they should be shared between many target units. It seems that it is really important to propose another method, yielding more accurate results on the district mesh. So, we propose an experiment based on the Modified Areal Weighting method using control zones.

3.2. Using Building Footprints as Control Zones

If we assume that weighted estimates can improve the reliability of the statistics, we propose to create a population potential for each residency building. Figure 4 illustrates the process for transferring IRIS population data towards Grenoble districts.


IRIS
Districts
Areal weighting disaggregation of population based on buildings as control zones
Aggregation of data
Buildings

$$pop_B = pop_{IRIS} \times \frac{Sdev_B}{\sum_{IRIS} Sdev_B}$$

$$pop_D = \sum (pop_{B \subset D})$$

Fig 4. Process for transferring population data from IRIS towards districts in Grenoble.

Using building elevation and floor height, depending on the age of the buildings, we can estimate the number of floors for each one, and compute the developed area

$Sdev_B$ (number of floors multiplied by footprint area). This is used to weight the disaggregation of population accounts of IRIS (pop_{IRIS}) onto buildings objects, and determine the number of inhabitants per building (pop_B). Then, we can sum population of buildings included inside each district, and get an estimate of population (pop_D) by district.

3.3 Analysis of the Results

Figure 5 shows a map of population accounts on the district mesh, obtained either by the simple areal weighting disaggregation, or by the modified areal weighting disaggregation, using buildings as control zones.

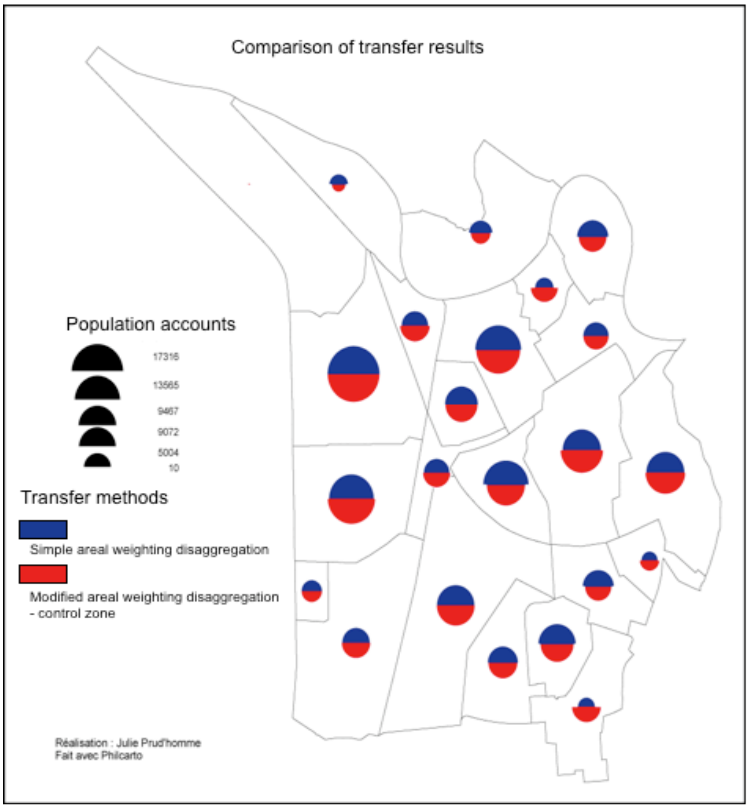

Comparison of transfer results
Population accounts
17316
13565
9467
9072
5004
10
Transfer methods
Simple areal weighting disaggregation
Modified areal weighting disaggregation - control zone
Réalisation : Julie Prud'homme
Fait avec Philcarto

Fig. 1. Comparison of population estimations in Grenoble: simple areal weighting (upper half-disk) versus modified areal weighting disaggregation (lower half-disk).

It can be noticed that both methods provide very similar results on a sub-set of districts, while, on other districts, a significant difference is measured (up to 3500 inhabitants for the biggest difference). The *Capuche* district, which is a typical case of a district which intersects many IRIS units, is also one of the zones presenting a very significant difference between both methods (almost 3800 inhabitants of difference).

In the *Villeneuve_VO* district, a significant difference (1106 inhabitants) can be noticed as well, but in favor of the modified areal weighting disaggregation, based on buildings as control zones. This was very likely to happen since this unit has a high density of buildings.

In order to assess estimates, we have at our disposal data provided by the French National Institute of Socio-Economic Studies (INSEE in French) at district level, and obtained by a census relevant only on one district, *Villeneuve_VO* and especially on *La Bruyère* area.. Indeed, *La Bruyère* is an area which is in *Villeneuve* district but not in IRIS *Villeneuve_VO*. Thus, the difference between them is used as a reference for the comparison of the results. This area includes miscellaneous types of housing (large group, mean group, individual [5]).

Table 1. Comparison of population estimations in *La Bruyère* area with disaggregation methods.

census data on *La Bruyère* area	Simple areal weighting disaggregation	Modified areal weighting disaggregation
458	602	1107

The comparison of our results with those data (Table 1) shows that the results based on simple areal weighting disaggregation method are closer to reality (which is 458 inhabitants in *La Bruyère* area). Concretely, these inhabitants are not recorded in the district to where they belong (equipments, policies etc.).

This result can be explained by a deeper study, which shows that footprint of building is in this case not a good ancillary data for computing the developed area in this district, having modern architecture. Indeed, the spatial images processing, that provides the building footprint, takes into account the numerous terraces, balustrades, recess or projections of buildings in this district. This increases artificially the extent of their footprint. This method has an underlying hypothesis of constant population density between buildings, which is wrong. In this example, housing accounts per building may have been a better ancillary data.

As a conclusion, this study case reveals that a transfer based on modified areal weighting with control zones is not always closer to census data than a simple areal weighting method. This result was not expected, but it corroborates other studies linked to the usage of land use data for disaggregation [32]. This confirms that data issued from spatial image processing cannot automatically replace data coming from census.

4 Conclusions

As an answer to the so-called "change of support problem", this work provides a first scheme for transferring variables between misaligned meshes. We have presented a state of the art of the transfer methods, organized into three categories: aggregation, disaggregation, and interpolation methods. Through this review, the main characteristics for modeling these methods have been highlighted: we distinguish the nature of the data support (punctual, regular polygonal, irregular polygonal), the kind of data (quantitative, or qualitative, absolute or ratio), and the scale of the support (finer or larger). Also, some parameters are common to certain families of methods (such as kernel smoothing or focal functions).

The accuracy of the results relies both on the method used, and on the quality of data. Indeed, the method, by its own internal mechanisms, will transform data (scale and zoning effects). But the quality (accuracy, fitness for use, etc.) of data involved into the transfer process (especially the ancillary data) has also a great influence on the quality of the result. A case study achieved on the city of Grenoble illustrates the difficulties that can be encountered. This outlines the very exploratory aspect (trials and error) of the transfer process. Further work could use some descriptive modeling methods (local or global autocorrelation indexes), and metadata usage in order to select the most appropriate ancillary data.

There are numerous software and libraries providing methods for transfer data, but there is no system integrating these various methods, connected to a spatial database, and offering some kind of assisted workflow for data transferring process between supports. This preliminary study will be used to design a task model (an object-oriented model describing a hierarchy of transfer methods, together with their conditions of use). This task model could be at the core of a system controlled by the user. For instance, the user could interact to establish the strategy for the task activation, and give some inputs linked to his/her own knowledge and expertise about data. This system should allow for estimation of missing data, or combination of heterogeneous data, linked with various territorial zonings.

References

1.Arbia G. “Statistical Effect of Data Transformations: A Proposed General Framework,” in The Accuracy of Spatial Data Bases, eds. M. Goodchild and S. Gopal, London: Taylor and Francis, (1989) 249–259.

2.Arnaud, M., Emery, X., Estimation et interpolation spatiale : méthodes déterministes et méthodes géostatiques, Paris, Hermès, (2000), (In French)

3.Badran, F., Daigremont, P., and Thiria S., « Régression par carte topologique »,. In: Statistiques et méthodes neuronales, eds. S. Thiria et al., (1997), 207-222. (In French)

4.Bracken, I., Martin, D. The generation of spatial population distributions from census centroid data. Environment and Planning A, (1989), 21: 537–543.

5.CERTU, Méthodes d’estimations de population, Lyon, (2005), (In French)

6.Cressie N., Statistics for spatial data, John Wiley and Sons, New-York (1991)

7.Droesbeke, J-J., Lejeune, M., Saporta, G., Analyse statistique des données spatiales, Paris, Technip, (2006), (In French)

8.Dubrule, O. « Two methods with differents objectives : splines and kriging ». Mathematical geology. (1983) 15: 245-255.

9.ESPON 3.4.3, The modifiable areas unit problem, Luxembourg, Final report, (2006)

10. EUROSTAT European Commission Statistical Office – EUROSTAT, GIS Application Development, Final Report, (1999)

11. Fisher P.F. and Langford M., Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation, in Environment and Planning A, (1995), 27: 211-224.

12. Flowerdew, R., Green, M. Statistical methods for inference between incompatible zonal systems, in M. Goodchild & S. Gopal, eds, 'The accuracy of spatial data bases', Taylor and Francis, London, (1989), 239–247.
13. Flowerdew, R., Green, M. "Developments in areal interpolation methods and GIS", in: The Annals of Regional Science, (1992) 26: 76-95.
14. Fotheringham, A.S., Brunsdon, C., and Charlton M., Quantitative Geography, Sage, London, (2000), 59-60
15. Gomez O, Paramo F., The Land and Ecosystem Accounting (LEAC) methodology guidebook, Internal report, (2005),http://dataservice.eea.europa.eu/download.asp?id=15490
16. Goodchild, M.F., Anselin, L., Diechmann, U., « A general framework for the areal interpolation of socio-economic data », in Environment and Planning A, (1993), 383-397
17. Gotway, C., Young, L, « Combining incompatible spatial data », in Journal of the American Statistical Association, (2002), 97(458): 632-648
18. Grasland, C., « A la recherche d'un cadre théorique et méthodologique pour l'étude des maillages territoriaux » Entretiens Jacques Cartier, « Les découpages du territoire », Lyon, décembre (2007). (In French)
19. Langford, M., Unwin, D.J., « Generating and mapping population density surface within a GIS », in The Cartographic Journal, (1992), 31: 21-26
20. Marceau, D, « The scale issue in social and natural sciences », in Canadian Journal of Remote Sens, (1999), 25(4): 347-356
21. Markoff, J., and Shapiro, G. "The Linkage of Data Describing Overlapping Geographical Units," Historical Methods Newsletter, (1973), 7: 34–46.
22. Matheron, G., "Principles of geostatistics", Economy geology, (1963), 58: 1246-1266.
23. Miller, H.,J., "Geographic representation in spatial analysis". Journal of Geographical Systems (2000) 2(1): 55-60
24. Nordhaus, W.D., « Alternative approaches to spatial rescaling », Yale University, New Haven, CT., (2002)
25. Openshaw, S., and Taylor, P., "A Million or so Correlation Coefficients," in Statistical Methods in the Spatial Sciences, ed. N. Wrigley, London: Pion, (1979) 127–144
26. Openshaw, S., « Building an automated modelling system to explore a universe of spatial interaction models », in Geographical Analysis, (1988), 20: 31-46
27. Plumejeaud, C., Vincent, J-M, Grasland, C., Bimonte, S., Mathian, H., Guelton, S., Boulier, J., Gensel, J. "HyperSmooth, a system for Interactive Spatial Analysis via Potential Maps", - In W2GIS 2008, Shanghai, December 11-12, China, (2008).
28. Rase W.D., « Volume-preserving interpolation of a smooth surface from polygon-related data », in Journal of Geographical Systems, (2001), 3: 199-213
29. Raskin, R.G., "Spatial Analysis on a Sphere: A Review". National Center for Geographic Information and Analysis, Technical Report (1994)
30. Reibel, M., Agrawal, A., "Areal Interpolation of Population Counts Using Pre-classified Land Cover Data," Population Research and Policy Review, Springer, (2007) 619-633
31. Rigaux, P., Scholl, M., « Multi-scale partitions: application to spatial and statistical databases », in SSD (1995), 170–183.
32. Schmit, C., Rounsevell M.D.A., La Jeunesse, I., The limitations of spatial land use data in environmental analysis, Environmental Science & Policy, (April 2006), 9(2): 174-188.
33. Shepard D. "A two-dimensional interpolation function for irregularly spaced data". Proc.23rd Nat. Conf. ACM 517–523 Brandon/Systems Press Inc., Princeton (1968)
34. Tobler, W. A., 'Smooth pycnopylactic interpolation for geographical regions', Journal of the American Statistical Association (1979) 74: 519–530.
35. Zaninetti, J.M., Statistique spatiale, méthodes et applications géomatiques, Paris, Hermès, (2005), (In French)
36. Zhang, Z., Griffith, D., "Developing user-friendly spatial statistical analysis modules for GIS: an example using ArcView", *Computer, Environment and Urban Systems*, (1993)