TÉCNICO LISBOA

Spatial Disaggregation Using Geo-Referenced Social Media Data as Ancillary Information

João Miguel Cordeiro Monteiro

Thesis to obtain the Master of Science Degree in

Telecommunications and Computer Engineering

Supervisors: Prof. Doutor Bruno Emanuel da Graça Martins
Prof. Doutor João Moura Pires

Examination Committee

Chairperson: Prof. Doutor Paulo Jorge Pires Ferreira
Supervisor: Prof. Doutor Bruno Emanuel da Graça Martins
Members of the Committee: Doutor Jacinto Paulo Simões Estima

October 2016

Acknowledgements

First, I would like to thank Professor Bruno Emanuel da Graça Martins and Professor João Moura Pires for their guidance during this last year, through which they contributed very significantly to the work that we have developed together, with their knowledge and motivation.

Second, I would also like to thank my family, in particular my parents and my sister, for their constant support and for giving me the opportunity to learn in such a distinguished institute as Instituto Superior Tećnico.

I would also like to thank Pierre Roudier, the author of an open-source software package that implemented the dissever algorithm in R with basis on original code from Malone et al. (2012), and with whom I have exchanged some information about spatial disaggregation in general. This software package was extended in the context of my work, for instance in order to support spatial disaggregation instead of just standard spatial downscaling, and in order to support geographically weighted regression models.

Finally, I have to thank all my friends and colleagues for the constant support during the hard, although also amazing, time spent at Instituto Superior Tećnico.

João Miguel Cordeiro Monteiro

For my parents and sister,

Resumo

A informação estatística relativa a atividades socioeconómicas da população, ou sobre aspetos de saúde pública, está amplamente disponível, apesar de a mesma ser apenas regularmente recolhida ou disponibilizada a um nível geoespacial relativamente agregado. A este nível agregado, diversas características sobre a informação tendem a ser mascaradas, por exemplo através da eliminação de *hotspots*, ou mesmo através da suavização de variações espaciais. Por estas razões, muitas vezes é necessário proceder a uma desagregação da informação, por forma a fornecer estimativas mais localizadas. No contexto de tarefas de análise espacial, desagregação ou *downscaling* espacial são técnicas que permitem transformar a informação disponível em zonas-origem (por exemplo distritos ou municípios) para outras zonas-alvo, com características geográficas diferentes e com uma maior resolução. Nesta dissertação é apresentada uma técnica de *downscaling*/desagregação, combinando procedimentos de análise de regressão do estado da arte com métodos clássicos de análise espacial, como sejam mapeamento dasimétrico ou interpolação picnofilática. O procedimento foi usado conjuntamente com informação auxiliar como densidade da população, estatísticas de tipo de ocupação do terreno, emissões de luzes noturnas, ou densidade de estradas no *OpenStreetMap*, para desagregar diferentes indicadores socioeconómicos. As experiências aqui reportadas envolveram ainda o uso de informação auxiliar georreferenciada extraída do Flickr, um conhecido serviço com base em localização, para produzir mapas de resolução fina respeitantes aos territórios de Portugal, Bélgica e França. É apresentada também uma discussão sobre a metodologia de desagregação espacial usada, e sobre a qualidade dos resultados obtidos em diferentes condições experimentais. Os resultados obtidos com a metodologia de desagregação referida superaram algoritmos seminais utilizados na literatura, como *mass-preserving areal weighting* ou interpolação picnofilática. Em adição, o procedimento produziu resultados consideravelmente melhores, quando utilizando a informação auxiliar proveniente da densidade de fotografias do Flickr, num estudo relativo à desagregação de estatísticas sobre turismo para o território da Bélgica.

Abstract

Statistical information on socio-economic activities or on public-health concerns is widely available, although the data are often collected or released only at a relatively aggregated level. Using aggregated data usually masks important local hotspots, and overall tends to over-smooth spatial variations in impact. For these reasons, we often need to disaggregate the source data, in order to provide more localized estimates. In the context of spatial analysis, spatial disaggregation or spatial downscaling are techniques that can be used to transform data from a set of source zones (e.g., districts or municipalities) into a set of target zones, with different geometry and with a higher general level of spatial resolution. This dissertation presents a hybrid spatial downscaling/disaggregation technique which combines state-of-the-art regression analysis procedures with the classic methods of dasymetric mapping and pycnophylactic interpolation. This procedure was used together with ancillary data, like population density, land coverage, nighttime light emissions, or OpenStreetMap road density, to disaggregate different types of socio-economic indicators. The reported experiments have also leveraged ancillary georeferenced data extracted from a popular location-based service, namely from Flickr, to produce high-resolution gridded datasets relative to the Portuguese, Belgian and French territories. A detailed discussion around the spatial disaggregation methodology and on the quality of the obtained results, under different experimental conditions, is also presented. The disaggregation results that were achieved outperformed those from seminal baseline algorithms that were previously used in the literature, such as mass-preserving areal weighting (Goodchild and Lam, 1980) or pycnophylactic interpolation (Tobler, 1979). Also, the procedure produced notably better results when leveraging the density of Flickr photos as ancillary information, in a study concerning with the disaggregation of tourism statistics in the territory of Belgium.

Palavras Chave
Keywords

Palavras Chave

Análise espacial

Sistemas de informação geográfica

Desagregação espacial baseada em regressão

Indicadores socioeconómicos

Informação georreferenciada de redes sociais

Keywords

Spatial analysis

Geographic information systems

Regression-based spatial disaggregation

Socio-economic indicators

Geo-referenced social media information

Contents

List of Figures

List of Tables

1 Introduction

Statistical information on socio-economic activities or on public-health concerns is widely available, although the data are often collected or released only at a relatively aggregated level. The main reasons why this aggregation occurs are: (1) spatial data concerning personal information are restricted by privacy and confidentiality; (2) data in the aggregated form requires less volume for storage; (3) geography has a long tradition of studying data at the regional level (Lin and Cromley, 2015b). However, using aggregated data usually masks important local hotspots, and overall tends to smooth out spatial variations in impact. For these reasons, we often need to disaggregate the source data, in order to provide more localized estimates. The major contribution of this work is the proposal of a novel method to spatially disaggregate socio-economic indicators. I considered several sources of ancillary information to aid in the procedure, such as population density or nighttime light emissions, as well as other types of ancillary knowledge that have been less studied in previous works, like geo-referenced social media data from services such as Flickr. To evaluate the proposed algorithm, I performed a large set of experiments concerning the geographic territories of Portugal, Belgium, and France, and compared the error values using state-of-the-art regression methods that combine the different sources of ancillary data, against simple baseline approaches.

1.1 Motivation

In the context of spatial analysis and geographic information systems, spatial disaggregation or spatial downscaling are processes by which information at a coarse spatial scale is translated to finer scales, while maintaining consistency with the original dataset. Spatial disaggregation techniques are thus used to convert data originally available for a set of source zones (e.g., districts or municipalities) into a set of target zones (e.g., raster cells with a resolution of 1x1 km), that have a different geometry and a higher general level of spatial resolution. While the term spatial disaggregation is most often used in the context of additive variables (i.e., population counts and other datasets of aggregated counts over which the pycnophylactic property should be enforced), spatial downscaling procedures are mostly focused on non-additive variables (e.g.,

different types of environmental or geophysical properties, such as temperature, precipitation, soil moisture, agricultural land usage, air quality, etc.).

From simple areal weighting to intelligent dasymetric mapping (Hawley and Moellering, 2005), most spatial disaggregation approaches have been applied to population data, and they have in common the mass-preserving, or pycnophylactic, property, in that the higher resolution estimates are conditioned to sum to the corresponding values in the source zones. Despite the fact that most previous studies concerning with spatial disaggregation/downscaling have focused either on population density (see for instance the WorldPop[1] project, described in the papers by Sorichetta et al. (2015) and Stevens et al. (2015)) or on geophysical/environmental variables, spatial disaggregation/downscaling can be equally useful in domains such as the social sciences and the analysis of socio-economic activities (Doll et al., 2000), public-health concerns (Vanwambeke et al., 2011), agriculture indicators (Chakir, 2009), or other economic activities (Matisziw et al., 2008). In fact, seminal work on the area by Goodchild et al. (1993) considered variables such as employment and income, and more recent work within the G-Econ research project of the University of Yale aimed to develop raster datasets on economic activity through spatial rescaling based on proportional allocation (Nordhaus, 2006, 2003).

Also referred to as a grid data model, the term raster denotes a type of tessellation (i.e., a mosaic) that is used to divide a geographic surface under study into uniform cells. The idea is to use this grid to represent the phenomenon under study over a geographic space. Through the usage of spatial disaggregation/downscaling to aid in the construction of raster datasets, it is latter easier to examine socio-economic patterns and link them to readily available geophysical data published in the form of grids, such as climate conditions, soil properties, land coverage and usage, ecology, and the like. Although socio-economic data is typically only available at the level of regional administrative units, the analysis of the information through different partitions of space can indeed be interesting to perform, in order to better investigate local hotspots, or to find relations to particular geophysical characteristics (e.g., proximity to regions with specific land coverage types, or relations towards terrain elevation).

1.2 Thesis Proposal

Given that location-based services are often used to share information relevant to a variety of socio-demographic issues, it would be interesting to implement a new and effective spatial

[1] `http://http://www.worldpop.org.uk`

disaggregation method for socio-demographic data published at a course granularity level, combining the ideas of dasymetric mapping and pycnophylactic interpolation, with the incorporation of ancillary data extracted from popular location-based services and social media sources, like Flickr[2]. Taking inspiration on very recent studies (Lin and Cromley, 2015a; Longley et al., 2015; Patel et al., 2016), I argue that this idea can for instance be implemented by creating density surfaces with basis on the number of geo-referenced items published on these services. The data from these location-based services, along with population density[3], nighttime light emissions[4] and OpenStreetMap[5] road density information, can provide very useful information to disaggregate different types of socio-economic indicators, published at a course granularity level (i.e., data published at the level of municipalities or civil parishes), in resources such as tables from the Portuguese National Institute of Statistics[6] or from the Eurostat portal[7]. The spatial disaggregation technique discussed in this thesis has, in fact, resulted in the production of raster datasets with different types of socio-economic indicators.

1.3 Contributions

In brief, the main contributions of this thesis are as follows:

- The comparison of different methods in the disaggregation of socio-economic indicators, including simple baseline approaches (e.g., mass-preserving areal weighting or pycnophylactic interpolation) and dasymetric mapping methods that leverage ancillary sources of information. Most previous work in the area has instead considered applications such as population mapping, and it would be interesting to see (i) if similar approaches can also be used in the disaggregation of other types of variables, and (ii) the degree to which different types of ancillary data can be used to improve the disaggregation results, depending on the type of variable that is being analyzed;

- The proposal of a novel intelligent disaggregation method, based on a downscaling procedure originally outlined by Malone et al. (2012), that uses regression analysis to combine different ancillary variables. This dissertation describes the adaptation/extension of the

[2] http://www.flickr.com
[3] http://sedac.ciesin.columbia.edu/data/collection/gpw-v3
[4] http://ngdc.noaa.gov/eog/viirs/download_monthly.html
[5] http://fred.dev.openstreetmap.org/density
[6] http://www.ine.pt
[7] http://ec.europa.eu/eurostat

original procedure by Malone et al. (2012) (e.g., combining it with the use of pycnophylactic interpolation) and its evaluation on the disaggregation of several indicators, when using different types of regression algorithms;

- A detailed evaluation of the proposed spatial disaggregation procedure, by conducting experiments using different regression methods to combine multiple sources of ancillary information. In particular, the results reported in this dissertation suggest that state-of-the-art regression methods, such as geographically weighted regression or robust regression, can indeed be useful in this application context, especially in the case of large study regions, given that the ancillary sources of information can violate the assumptions of standard linear regression, and given that the optimal coefficients that correspond to the weights of each ancillary variable can also vary significantly over the geographic space;

- The realization of experiments with other types of ancillary data that have been less studied in the literature (i.e., social media data extracted from location-based services, like Flickr), to aid in spatial disaggregation procedures. For instance, the usage of these types of ancillary information, through the creation of density surfaces based on the occurrence of geo-referenced items, was shown to result on notably better results (i.e., lower error metrics) in the application of the procedure to particular geographic regions, in a study concerning with the disaggregation of tourism statistics. This improvement is even more evident when the target zones correspond to regions of high activity in Flickr, since this will produce a more reliable density map of social media usage, and then a more accurate correlation with the indicators under study.

Part of the work reported on this dissertation was also presented on an article entitled *Spatial Disaggregation of Socio-Economic Indicators*, that was submitted to the Geographical Analysis Journal from Wiley. The article reports on spatial disaggregation experiments specifically leveraging data relative to the Portuguese territory, resulting in the production of raster datasets with different types of socio-economic indicators. The datasets used in this case study relied on information from the year of 2011 (i.e., the year of the last national census study), and focused on several themes like population, environment, or tourism, among others. The disaggregation method consists of three main consecutive logical steps, namely (i) a simple iterative pycnophylactic-interpolation process for a preliminary data redistribution, (ii) a proportional and weighted areal interpolation using population as ancillary data, and (iii) a disseveration-based procedure to combine both previous estimates with other ancillary variables. The disaggregation method that was extended/adapted from the original procedure outlined by Malone

et al. (2012), and that is discussed in this dissertation and in the aforementioned article, is now also publicly available as an open source software package[8].

1.4 Structure of the Document

The rest of this document is organized as follows. Chapter 2 presents fundamental concepts and seminal geospatial analysis methods, as well as important related work in the areas of spatial downscaling/disaggregation. Then, Chapter 3 describes the different steps involved in the considered spatial disaggregation approach, detailing also the ancillary datasets and the different types of regression methods that were employed in the experiments. Chapter 4 presents the results of evaluation experiments, with different case studies relative to the Portuguese, Belgian, and French territories. Finally, Chapter 5 concludes this document by summarizing the main findings of this work, and highlighting possible directions for future research.

[8] `http://github.com/bgmartins/dissever`

2 Concepts and Related Work

This chapter presents fundamental concepts and related work on spatial disaggregation. First, an overview on seminal methods is provided in Section 2.1. This overview consists in presenting the most relevant methods for spatial data disaggregation in Section 2.1.1, and other spatial data analysis methods in Section 2.1.2. Then, important related work in the area is presented in Section 2.2, covering five main categories: Simple areal interpolation (Section 2.2.1), approaches leveraging Earth observation products as ancillary data (Section 2.2.2), approaches leveraging Earth observation data together with other sources of ancillary data (Section 2.2.3), approaches leveraging mobile phone call records and other types of point-based ancillary information (Section 2.2.4), and approaches leveraging social media as ancillary information (Section 2.2.5). Finally, an overview on the related work is presented in Section 2.2.6.

2.1 Fundamental Concepts

Spatial disaggregation, as a procedure, is applied to data sets for which the real underlying spatial distribution is unknown, but for which aggregated data already exist. The process of spatial disaggregation thus corresponds to the transformation of data from the arbitrary zones of data aggregation to a set of target zones with different geometry and a higher general level of spatial resolution. The increased development of Geographic Information Systems (GIS) has also increased the necessity for spatial disaggregation, as a GIS analysis frequently generates new layers of areal units for which non-spatial attribute information must be estimated (Lin and Cromley, 2015b). Some fundamental concepts regarding spatial disaggregation procedures are presented next, followed by a short introduction to other types of elementary spatial data analysis operations, such as kernel density estimation or kriging.

2.1.1 Spatial Data Disaggregation Methods

This section overviews seminal methods for spatial data disaggregation, including mass-preserving areal weighting, binary dasymetric mapping, general dasymetric mapping, and pyc-

nophylactic interpolation.

2.1.1.1 Mass-Preserving Areal Weighting

The mass-preserving areal weighting method is the simplest algorithm for performing spatial disaggregation and it assumes a homogeneous distribution of the data throughout each source zone (Goodchild and Lam (1980) and Goodchild et al. (1993)). Despite being very straightforward, it is a volume-preserving disaggregation algorithm that conserves the total value within each zone, and therefore no adjustment procedure is needed over the results of the original procedure, in order to ensure the pycnophylactic property. It can also be easily implemented in both vector and raster GIS environments.

Given a variable to be disaggregated, the total value for a target zone is estimated by performing a weighted summation of the density values of all source zones falling in the target zone, according to the following equation:

$$P_t = \sum_{\{s:s\cap t\neq\emptyset\}} (\alpha_{s,t} \times P_s) \tag{2.1}$$

In Equation 2.1, P_t is the estimated count in target zone t, while P_s is the count in source zone s. The parameter $\alpha_{s,t}$ is computed using the overlapping area between the source zone and the target zone, being used as a weight.

$$\alpha_{s,t} = \frac{A_{ts}}{A_s} \tag{2.2}$$

In Equation 2.2, A_s corresponds to the area of the source zone s, and A_{ts} is the area of target zone t overlapping source zone s.

For example, in the scenario illustrated by Figure 1, representing the disaggregation of population using mass-preserving areal weighting, S1, S2, S3 and S4 are the source zones, for which the respective values of population are known. The region G corresponds to a grid with the target zones, and C is one of the cells for which we want to get the estimated value of population. Since the area of cell C that overlaps the source zone S3 is approximately of 18 measurement units (m.u.) and the area of the source zone S3 is approximately 324 m.u., the estimated value using the method from Equations 1 and 2 is equal to $P_c = \frac{18}{324} \times 30 = 1.67$.

Mass-preserving areal weighting is perhaps the most commonly used spatial disaggregation

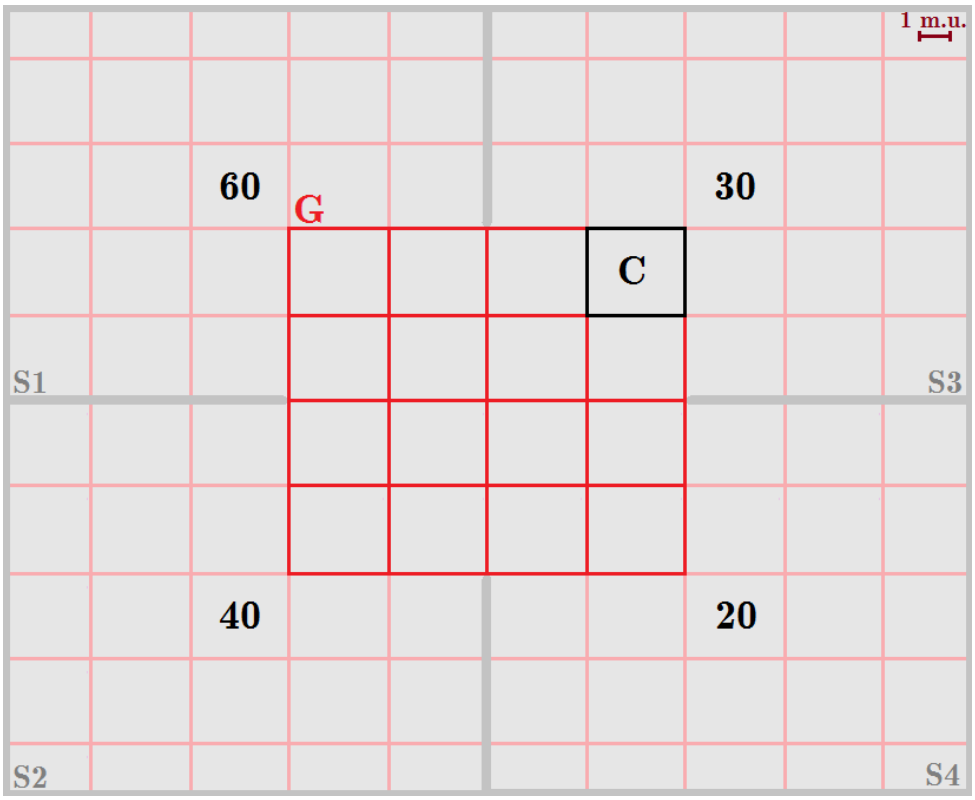

Figure 2.1: Spatial disaggregation using mass-preserving areal weighting.

method, due to its intuitively simple theory, as well as its low data and computation requirements. This method is probably also the only choice when no additional information is available for the interpolation area. It should nonetheless be noted that areal weighting is based on the often incorrect assumption that the phenomena of interest is evenly distributed across the source zones. For instance, most populations are rarely uniform across census tracts, and instead tend to be highly clustered in urban centers, surrounded by areas of dispersed rural homesteads. This method also produces poor estimates when compared against the results of other methods, because of its simplistic representation of the density surface, and due to the fact that it does not use any sources of ancillary data.

2.1.1.2 Binary Dasymetric Mapping

Binary dasymetric mapping, also referred to as masked areal weighting, is a direct extension of areal weighting that uses ancillary data to build a mask that defines where, within the target zones, the source data should be allocated (Eicher and Brewer, 2001). In fact, areal weighting is used to estimate target zone populations from the source zones based on populated areas, and thus binary dasymetric mapping can also be termed as an areal weighting of populated areas. The ancillary data is used to classify the target space into areas containing population and

areas that do not, and therefore this approach assumes that the variable might not be evenly distributed in the source zone, although it is evenly distributed in the control zones within the source zone. The general procedure also corresponds to the computation shown in Equation 2.1, with weights $\alpha_{s,t}$ computed as follows:

$$\alpha_{s,t} = \frac{A_{ts}^{(p)}}{A_{s}^{(p)}} \tag{2.3}$$

In Equation 2.3, $A_{ts}^{(p)}$ is the area of populated land that overlaps between the target map unit t and source map unit s, while $A_{s}^{(p)}$ is the area of the target map unit s that corresponds to populated land.

For example, in the scenario illustrated by Figure 2, representing the disaggregation of population using binary dasymetric mapping, S1, S2, S3 and S4 are the source zones, while G is the grid corresponding to the target zones. The region C is one of the cells for which we want to get the estimated value of population, and M1, M2, M3 and M4 are the masks that define where the population should be allocated within each source zone (i.e., the populated zones). Since the area of cell C that overlaps the mask M3 is approximately 4 m.u., and since the area of the mask M3 is approximately 16 m.u., the estimated value using the weights $\alpha_{s,t}$ from Equation 2.3 is equal to $P_c = \frac{4}{16} \times 30 = 7.5$.

Binary dasymetric mapping is also easy to implement in GIS. In fact, it has gained popularity with the increasing availability of satellite imagery and with improved methods for leveraging Earth observation data, since these can be used do identify populated areas and create the mask. Therefore, the results of binary dasymetric mapping are generally an improvement over simple areal weighting. However, despite the benefits, there are still considerable deficiencies in this method. Binary dasymetric mapping implicitly assumes that population density is constant in all the populated areas within a source zone, which is a weakness since this does not often occur in reality. For instance, populated areas usually do not have the same density, although binary dasymetric mapping considers that. Additionally, non-populated areas, that often have some population, are completely eliminated from the total area of the source zone. Several authors tried to evolve this approach from a binary model to more nuanced ones, which result in a more realistic depiction of densities encountered in the real world.

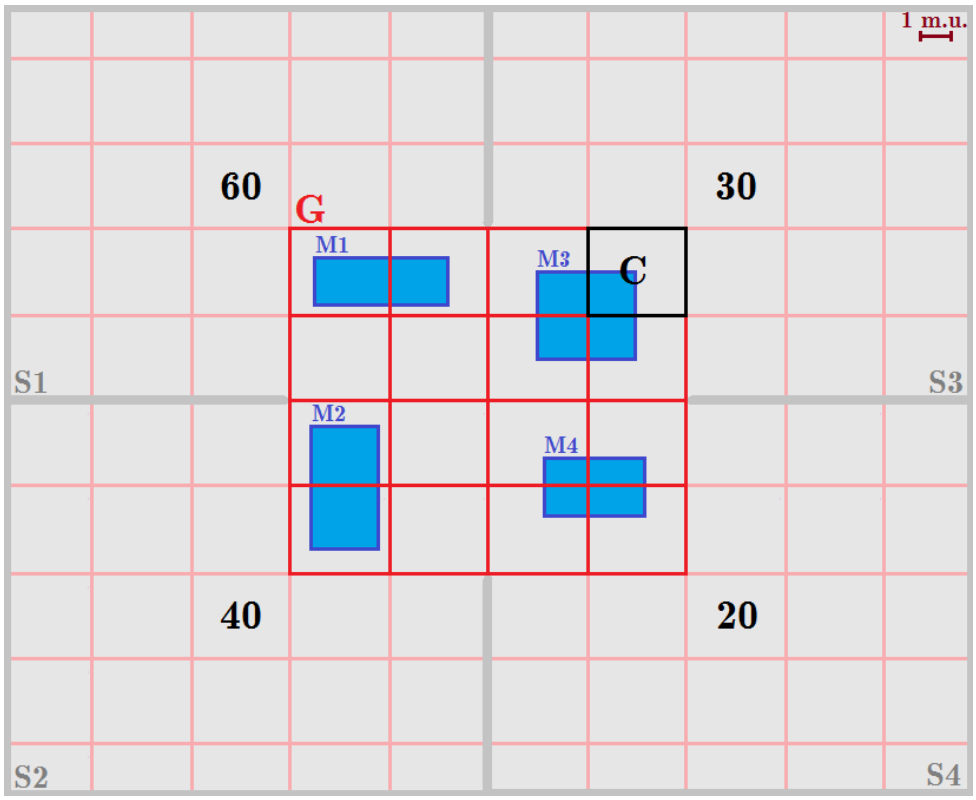

Figure 2.2: Spatial disaggregation using binary dasymetric mapping.

2.1.1.3 General Dasymetric Mapping

General dasymetric mapping, also referred to as polycategorical dasymetric disaggregation or as the class percent method, is an improvement over binary dasymetric mapping in that two or more categories can be assigned weights for disaggregation, for example to reflect different levels of population density, rather than assuming that population is evenly distributed in all residential units in a source (Eicher and Brewer (2001); Mennis and Hultgren (2006)). This method incorporates ancillary data to facilitate the interpolation process and it applies percentages to each of the categories for the source area, representing the percentage of the variable that is likely to be contained within that category, per source area.

In an extreme case, where we do not have a predefined number of classes, this disaggregation method can be seen as a proportional and weighted areal interpolation approach, where each target zone also takes a proportionally calculated value, but in this case the value is weighted according to some external variable. The general method corresponds to the procedure shown in Equation 2.1, with weights $\alpha_{s,t}$ computed as follows:

$$\alpha_{s,t} = \frac{W_t \times A_{ts}}{\sum_{\{t':t' \cap s \neq \emptyset\}} W_{t'} \times A_{t'}} \tag{2.4}$$

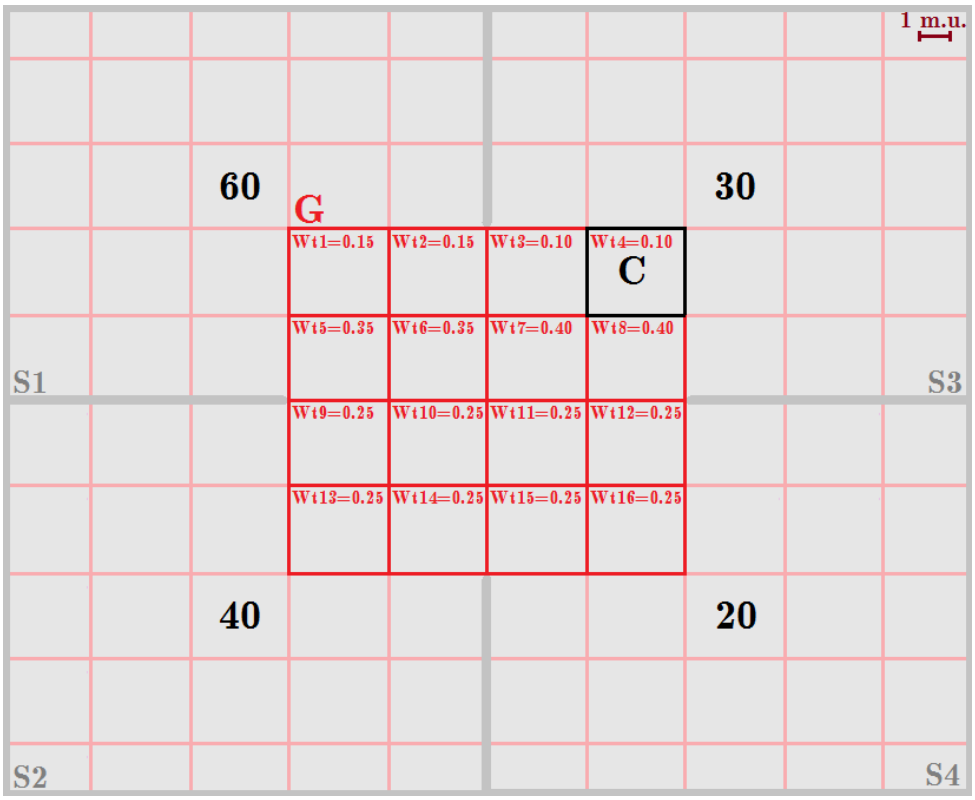

Figure 2.3: Spatial disaggregation using general dasymetric mapping.

In Equation 2.4, W_t is the weight assigned to the target zone t, and each W_t is chosen with basis on an external variable (or a combination of external variables), ensuring that $\sum_{\{t':t'\cap s\neq\emptyset\}} W_{t'} = 1$ if the estimates are required to sum to the same values of the source zones.

For example, in the scenario illustrated by Figure 3, representing the disaggregation of population using general dasymetric mapping, S1, S2, S3 and S4 are the source zones for which the respective values of population are known. The area G is the grid corresponding to the target zones, and C is one of the cells for which we want to get the estimated value of population. The target zones represented in the grid have weights, as indicated in the figure, that correspond to the percentage of population of the source zone that is contained in the target zone. Since the area of each cell that overlaps the source zone S3 is approximately 18 m.u., the estimated value using the weights $\alpha_{s,t}$ from Equation 2.4 is equal to $P_c = \frac{0.10\times18}{0.10\times18+0.10\times18+0.40\times18+0.40\times18} \times 30$, which in turn is equal to $P_c = \frac{1.8}{18} \times 30 = 3$.

The main challenge in dasymetric disaggregation involves finding an appropriate set of weights that can be applied to the land parcels, on the basis of ancillary data, to accurately reflect population density. These weights may, for instance, be defined using selective sampling or by some form of regression analysis. Most modern disaggregation methods have actually evolved from this approach.

2.1.1.4 Pycnophylactic Interpolation

Pycnophylactic interpolation was proposed by Tobler (1979) for isopleth mapping, and was applied to areal interpolation by Goodchild and Lam (1980). It is an extension of simple areal weighting that assumes a degree of spatial auto-correlation for the variable being interpolated.

The technique first computes the values for the target zones, according to Equations 1 and 2, similarly to the areal weighting method. This density, however, is then smoothed over each cell to combine the impacts of the adjacent neighbors on its grid value. The value of a cell is often computed as the average of its four orthogonal neighbors or all its immediate eight neighbors in the 3-by-3 smoothing window. The size of the window can be customized based on the characteristics of the underlying data. The predicted values in each source zone are then compared with the actual values and adjusted (e.g., if the estimated value of the source zone is 5% lower than the actual value, the values of each cell in that source zone are increased by 5%) to meet the pycnophylactic property of mass-preservation, with an iterative procedure that continues until there is either no significant difference between the estimated values and the true values within the source zones, or until there have been no significant changes on the cell values from the previous iteration. To avoid multiple iterations of calculation and save computation time, more advanced procedures are often employed in practice, to adjust the smoothed values (Zhang and Qiu, 2011).

The interpolated surface produced by the pycnophylactic method is smooth, with relatively small changes in attribute values at target region boundaries. The sum of combined target attribute values, within each source region, is also kept consistent. However, by creating this smoothed surface, pycnophylactic interpolation does not confine itself to the homogeneity assumption of the areal weighting approach. It is an improvement over the area-weighting method, but makes no attempt to use the abundant sources of information that are available and that can be considered as ancillary data. In addition, this method only works within a raster GIS environment, since it requires a raster smoothing process, as well as vector-to-raster and raster-to-vector operations as intermediary steps. Practical implementations[1] are nowadays available within the R[2] system for statistical computing.

[1] `http://cran.r-project.org/web/packages/pycno/index.html`
[2] `http://www.r-project.org`

2.1.2 Other Spatial Data Analysis Methods

Other types of spatial data analysis operations, commonly used as pre- or post-processing in the context of spatial disaggregation, are presented next.

2.1.2.1 Kernel Density Estimation

Kernel density estimation is a non-parametric procedure to obtain the probability density function of a random variable, e.g., a variable whose value changes over the geographic space. In comparison to parametric estimators, where there is a fixed functional structure, and where the parameters of the function are the only information needed, non-parametric estimators have no fixed structure and depend upon all the data points to reach an estimate (Hwang et al., 1994).

Kernel estimators center a kernel function at each data point. In other words, they smooth the value of that data point with a contribution from each observed data point over a local neighbourhood. The contribution of a data point x_i to the estimate at some point x' depends on how apart x_i and x' are. The extent of this contribution is dependent on the kernel function being adopted, and on a bandwidth h. The estimated density at any point x', using the kernel function K() and considering a set of N observed data points, is given by:

$$\hat{\mathrm{f}}(x') = \frac{1}{N} \sum_{i=1}^{N} \mathrm{K}\left(\frac{\mathrm{d}(x', x_i)}{h}\right) \tag{2.5}$$

In Equation 2.5, $\int \mathrm{K}(t)dt = 1$ to ensure that the estimates $\hat{\mathrm{f}}(x')$ sum to 1, and d() is a given distance function from a known point x_i to the unknown point x'. The kernel function K() is usually chosen to be a smooth unimodal function with a peak at zero. Even though Gaussian kernels are the most often used, there are various possible choices for kernel functions.

Previous studies have shown that the quality of a kernel density estimate depends less on the shape of the kernel function than on the value of its bandwidth h. The choice of the most appropriate bandwidth is crucial, since a value that is too small or too large will lead to poor results. For instance, small values of h lead to very spiky estimates (i.e., not much smoothing), while larger h values lead to oversmoothing. Some previous studies have focused on the selection of an appropriate bandwidth to be employed in kernel density estimation procedures, for instance by using pilot estimation of derivatives (Sheather and Jones, 1991), by defining the bandwidth according to the minimum of the standard deviation (Scott, 1992), or through empirical cross-validation procedures (Venables and Ripley, 2002).

2.1.2.2 Inverse Distance Weighting Interpolation

Inverse distance weighting (IDW) is a deterministic and nonlinear interpolation technique, that uses a weighted average of the phenomenon values, from nearby sample points, to estimate the value of that attribute at non-sampled locations. The weight of a particular sampled point is computed depending on the sampled point's distance to the non-sampled location. The method is called inverse distance weighting because, according to Tobler's first law of geography, the similarity of two locations decreases with the increasing of the distance between them. A simple IDW weighting function, as defined by Shepard (1968), is as follows:

$$\mathrm{W}(x', x_c) = \frac{1}{\mathrm{d}(x', x_c)^p} \tag{2.6}$$

In Equation 2.6, x' denotes an interpolated arbitrary point, while x_c is an interpolating known point (i.e., a control location), and d() is a given distance function from the known point x_c to the unknown point x'. The parameter p is a positive real number, called the power parameter. The predicted value $\hat{\mathrm{f}}(x')$ for an unmeasured location x' is then given by:

$$\hat{\mathrm{f}}(x') = \frac{\sum_{i=1}^{N} \mathrm{W}(x', x_i)\mathrm{f}(x_i)}{\sum_{i=1}^{N} \mathrm{W}(x', x_i)} \tag{2.7}$$

In Equation 2.7, $\mathrm{f}(x_i)$ is the measured value at the i-th location, while $\mathrm{W}(x', x_i)$ is the weight for the measured value at the i-th location. The parameter x' is the prediction location, and N is the number of the surrounding measured values. Since inverse distance weighting is a deterministic technique, it does not take into account the spatial structure of the sample points. Therefore, the results obtained using this technique can be influenced by the density of the samples (e.g., a region with a higher density of samples may imply a more accurate estimate). Also, because inverse distance weighting computes an average value, the value for a non-sampled point can never be higher than the maximum value for a sample point, or lower than the minimum value of the sample point. Therefore, if the peaks of the data are not represented in the sample, this technique will be inaccurate in some locations.

2.1.2.3 Kriging

Kriging, or Gaussian process regression, is a method of interpolation in which the estimated values are modeled by a Gaussian process, assuming that the distance between sample points reflects a spatial correlation that can be used to explain the variation in the estimated geographic

region (Burrough and McDonnell, 1998). The surrounding measured values are weighted to derive a predicted value for an unmeasured location, using the following formula:

$$\hat{\mathrm{f}}(x') = \sum_{i=1}^{N} \mathrm{W}(x', x_i)\mathrm{f}(x_i) \tag{2.8}$$

In Equation 2.8, $\mathrm{f}(x_i)$, $\mathrm{W}(x', x_i)$, x' and N have the same meaning as in Equation 2.7. Unlike inverse distance weighted interpolation, the weights $\mathrm{W}(x', x_i)$ are based not only on the distance between the measured points and the prediction locations, but also on the overall spatial variation among the measured points. In simple kriging, the weights $\mathrm{W}(x', x_i)$ are given by the following equation system:

$$\begin{bmatrix} \mathrm{W}(x', x_1) \\ \dots \\ \mathrm{W}(x', x_n) \end{bmatrix} = \begin{bmatrix} \mathrm{c}(x_1, x_1) & \dots & \mathrm{c}(x_1, x_n) \\ \dots & \dots & \dots \\ \mathrm{c}(x_n, x_1) & \dots & \mathrm{c}(x_n, x_n) \end{bmatrix}^{-1} \times \begin{bmatrix} \mathrm{c}(x_1, x_0) \\ \dots \\ \mathrm{c}(x_n, x_0) \end{bmatrix} \tag{2.9}$$

In Equation 2.9, each $\mathrm{c}(x, y)$ corresponds to the known covariance between x and y. It is important to notice that, besides simple kriging, there are two more main kriging variants, specifically ordinary kriging, and kriging with a trend. While simple kriging is based on the assumption that the mean is constant over the entire domain, in ordinary kriging this mean is assumed to be constant only in the local neighborhood of each estimation point (i.e., for each nearby data value). Finally, kriging with a trend (also known as universal kriging) fits a linear or higher-order trend on the data points, instead of fitting just a local mean in the neighborhood of the estimation point.

Kriging is most appropriate (e.g., preferable over inverse distance weighting) when it is known that there is a spatially correlated distance or directional bias in the data. It also helps to compensate for the effects of data clustering, assigning individual points within a cluster less weight than isolated data points (Burrough and McDonnell, 1998). More details about kriging, and also about inverse distance weighting, are given in the books by Stein (1999) and by Burrough and McDonnell (1998).

2.1.2.4 Regression Analysis and Geographically Weighted Regression

Regression analysis is a statistical process for estimating the relationships between variables, including some techniques for modeling and analyzing them, when the focus is on the relationship

between a dependent variable and one or more independent variables. This way, regression helps in understanding how the value of the dependent variable changes with the variation of the independent ones. Regression analysis is also used to understand which independent variables are related to the dependent variable, and to explore the forms of these relations.

Many techniques for carrying out regression analysis have been developed. Familiar methods, such as standard linear regression, are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from data.

In linear regression, the dependent variable y_i is a linear combination of the parameters. For example, in a multiple linear regression for modeling N data points, there are several independent variables (e.g., covariates x_1 and x_2), as well as multiple parameters (e.g., β_0, β_1 and β_2) that correspond to the weights given to the different covariates. The goal is often to minimize the sum of the squares of the differences between the observations and the linear function that predicts missing data, taking into account that smaller differences indicate a better model for fitting the data. The equation for the aforementioned intuition is as follows, where ε_i represents an error term, and where p corresponds to the number of covariates:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon_i \tag{2.10}$$

The least squares method, applied to this specific model, requires the estimation of the values $\beta_j, j = 0, 1, ..., p$, with N measured values, so that $\sum_{i=0}^{N}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2$ is minimized. The values of β_j that minimize the sum of the squares of the residuals are computed by setting the derivatives of the aforementioned sum, with respect to each β_j, in turn equal to zero. Taking into account the $p + 1$ given normal equations that are solved simultaneously, the $p + 1$ least squares parameters are obtained. Since the problem becomes increasingly difficult as the number of independent variables increases, matrix algebra is often used to compute the regression results for more complex models.

Geographically weighted regression (GWR) is a particular variant of regression analysis, loosely based on kernel density estimation, in which the estimated parameters are location dependent. Taking into account that the fitted coefficient values of a global model may not represent detailed local variations in the data adequately, GWR calibrates multiple local regression models to examine the spatial variability of regression results across different regions, instead of computing one set of global regression values. These local results are produced by moving a weighted window over the data and by estimating one set of coefficient values at every chosen

fit point (Fotheringham et al., 1998). The GWR approach for estimating the regression parameters, using the least squares method, is an extension of the linear regression case. However, in GWR, one set of parameters is obtained for a set of local zones, determined by the weighting window, and given the specific local sample. If the local subset becomes too big, the error of the coefficient estimates is lower, but the chance that the coefficient deviation introduces bias is higher. To reduce this effect, a weighted least squares method is used.

In weighted GWR, the weights of the observations are not constant. Instead, these weights are computed according to the proximity to a control point, assigning a higher value to the observations that are close. Assuming that (u, v) are the coordinates of the the position of some data point in the study area, a typical GWR model that uses weighted least squares regression, in which the weights are a function of distance from location (u, v), is as follows:

$$y_i(u,v) = \beta_0(u,v) + \beta_1(u,v)x_1 + \beta_2(u,v)x_2 + \ldots + \beta_p(u,v)x_p + \varepsilon_i(u,v) \tag{2.11}$$

The distance between two points can be defined either by the actual geographic distance, or by a k-nearest neighbors algorithm. A bandwidth (i.e., a value that is typically associated to a Gaussian kernel) is selected to define the importance of the distance between sampled points and control points, which can be fixed or adaptative. Previous works showed that smaller bandwidths fit the data better but produce a high level of coefficient variability, and that fixed bandwidths often produce smoother spatial distributions than adaptive ones (Guo et al., 2008). Still, as discussed later in this dissertation, our experiments have leveraged adaptive bandwidths. As the bandwidth size increases, the GWR parameter estimates are more spatially similar to the parameter estimates obtained with standard linear regression.

When using GWR as a predictor (Harris et al., 2010; Perez-Verdin et al., 2014), the standard approach involves the production of leave-out-one predictions for the cases in which the training locations are also used for prediction, combined with the predictions at the remaining test data (i.e., when new locations are being used for prediction). The finer the spatial distribution of prediction points, the greater the computational burden, as separate regressions are calculated at each location. In practice, GWR is computationally very demanding.

2.2 Related Work

This section overviews important related work in the area of spatial disaggregation, also discussing previous studies specifically addressing the usage of geo-referenced social media data

for spatial disaggregation.

2.2.1 Simple Areal Interpolation

In a study of areal interpolation for socioeconomic data, Goodchild et al. (1993) developed a framework that addresses the problem of spatial analysis using non-coincident areal units. To achieve that, mass-preserving areal weighting was used, requiring the identification of one or more underlying continuous surfaces and assuming that the densities in the source zones were uniform. This way, it was possible to transfer attributes from one set of spatial objects to another, within a defined portion of the geographic space. The mass-preserving areal weighting procedure was already described in Section 2.1.

The authors illustrated the use of the framework with an example application for regions in the state of California, namely the 58 counties of California (i.e., the source zones) and the state's 12 major hydrological basins (i.e., the target zones). In order to conduct an economic impact study of water usage and policy, the socioeconomic data that was only available on the county level had to be transferred to the hydrological regions, for which the data connected with water issues were collected. The boundaries of the two datasets were, for the most part, incompatible, and thus a disaggregation method was required.

The case study presented by the authors was somewhat unrealistic and the results obtained using an areal weighting method revealed a much higher mean percentage error, when compared with other methods leveraging statistical approaches like pycnophylactic interpolation. To overcome the limitations found in their method, the authors suggested to incorporate some forms of ancillary knowledge into the disaggregation procedure.

2.2.2 Principled Approaches for Leveraging Earth Observation Products as Ancillary Data

In the context of one of the first attempts to obtain global maps of socioeconomic parameters and CO_2 emission, Doll et al. (2000) reported a dasymetric disaggregation approach using a time series of nighttime satellite images acquired between October 1994 and March 1995, combining this with other ancillary statistical information. A dataset was produced from information collected in cloud-free nighttime overpasses made by satellites of the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS), eliminating transient light sources such as migrating bush fires and shipping fleets. Other sources of ancillary data,

used to interpret the satellite data, were added through datasets of human population location centers, population densities, aggregated country-level statistics, and standard statistics on different aspects of the human environment.

The data was processed and analyzed within a GIS system, using a spatial intersection algorithm that assigns lit polygons to countries. Country-level relationships between lit area, gross domestic product, and total CO_2 emissions, were also used in a dasymetric mapping approach to create global maps of these parameters. The country-level relationships were discovered through simple linear regression approaches, where the lit area is the independent variable, while the gross domestic product and the total CO_2 emissions are the dependent ones.

The authors refer that the use of nighttime light emission data to map socioeconomic and CO_2 emission parameters has several advantages, such as the potential frequency of acquisition of the data and the differentiation of urban areas and transportation route locations, mapping quite well urban areas for global land-cover maps. However, they found some anomalies when building the CO_2 map, due to the exclusive use of the lit area as the parameter to derive relationships. A radiance value recorded at the sensor, modulating the magnitudes associated with each cell by acting as a weighting function, is referred to likely improve the ability of nighttime lights to map such parameters at the sub-national level.

Elvidge et al. (1997) reported on another dasymetric approach, but using different kinds of ancillary data. With the objective of modelling the spatial distribution of population density, carbon emissions, and economic activity, the authors estimated the area lit by visible-near infrared emissions for 21 countries, using also data from the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS). The authors concluded that these data are highly correlated to gross domestic product and electric power consumption.

The proposed method aimed to create a geo-referenced grid, reporting the normalized percent frequency with which lights were detected in cloud-free data. For that, time series of satellite observation data were used to obtain sufficient cloud-free observations to distinguish spatially stable lights from ephemeral events such as fire, lightning and system noise. The thermal band from OLS data collected in dark nights during 1994 and 1995 was used to identify clouds, while emission sources were identified by an algorithm that established a brightness threshold for lights based on the local background.

The results presented in the paper indicate that DMSP-OLS can be useful in several applications, e.g. to define and update the spatial distribution of human population on a global basis. For applications that require high spatial resolution, DMSP-OLS can also be used as a

guide in planning satellite data acquisitions. However, there is a huge regional influence on the relation between the area being lit and population numbers, and extra information to generate population maps from the DMSP stable lights should be required. These data can support the identification of urban areas, but DMSP stable lights cannot provide direct detection of rural population numbers across large land areas. Although this was not discussed by the authors, I believe that methods relying on geographic weighted regression can perhaps be more adequate for leveraging this type of information, given their ability to adjust parameters locally.

In the context of an EU-wide application, Briggs et al. (2007) developed a model that incorporates ancillary data, specifically EO nighttime light intensity data and CORINE land cover information, as covariates of population density in a GIS-based regression approach, disaggregating NUTS 5 census totals to a resolution of $1km^2$. In this particular approach, the steps to model population density are as follows. First, light emission data from DMSP satellites are sampled and modeled using kriging or inverse distance weighting (i.e., a range of models was tested and compared using the two methods, separately), to reduce the effects of blooming and provide a 200m resolution light emissions map. More details about kriging and inverse distance weighting are given in the books by Stein (1999) and by Burrough and McDonnell (1998), and Section 2.1.2 of this dissertation has also introduced these methods. The results are matched to land cover data, to enable more reliability in the distribution of the light emissions. Finally, local relationships are analyzed to derive weights to specific CORINE land cover parcels, used in the dasymetric disaggregation procedure to redistribute the population totals to each raster cell, according to the pycnophylactic property. The regression analysis and spatial interpolation techniques that were previously referenced (i.e., kriging and inverse distance weighting) were adopted to find the local relationships between variables, using the population count as the dependent variable and the area of lit land, area of unlit land, and total light emissions for each land region, as predictor variables. For inverse distance weighting, an inverse square function was applied as the distance function, with a search radius of 10km, a minimum number of sample points in each estimation of 4, and the maximum number of contributory points set to 100. For the case of kriging, several models were tested and compared, including linear, cubic and quadratic models, with and without global trends.

The authors argue that land cover, along with light emissions data, provides a useful indicator of where people live. By deriving weights for each land cover parcel that reflect population density, Briggs et al. (2007) built a high-resolution map of the European Union, dealing with common problems in the data: (1) census districts may change over time as a result of administrative restructuring; (2) surface reflection, scattering and refraction in the atmosphere can

cause ambiguity in the signal received by the satellite sensors; and (3) cloud cover and variations in the relationship between light emissions and population density from one area to another.

Although Briggs et al. (2007) presented a solution that is a significant improvement on simple area-weighting methods, some limitations are also recognized. Firstly, the data that was used does not conform fully to the demands of regression analysis, in that they often depart from conditional normality. The authors gave some considerations about this problem, in particular about normalising the data or using index values to transform the independent variables. Still, none of these methods was considered appropriate and, therefore, modeling was carried out with un-normalized variables. Although the authors have not discussed this possibility, I believe that robust regression methods (Huber et al., 1981) can can be an interesting alternative. Secondly, no account is also taken of spatial auto-correlation, mainly due to the fact that Bayesian techniques for addressing this issue were not computationally feasible, given the large datasets that were involved. Finally, there are also some uncertainties in the relationships between light emissions and population distribution or density, because light emissions depend on affluence and economic structure, at the national level, and on urban configuration, transport infrastructure, energy lighting policies, and lighting technology.

Gallego (2010) described four methods for disaggregating population totals from larger administrative units (i.e., NUTS 2 regions in Europe) to smaller ones (i.e., communes, with heterogeneous area, mostly between 10 and 100 km^2), combining population data per commune with land cover information. The land cover map used in this study is the 1-ha resolution raster version of CORINE Land Cover 2000 (CLC), while the population data was collected from LUCAS (Land Use/Cover Area frame Survey), i.e, a study that covered, in the year of 2001, the 15 member states of the European Union at that time.

The first method stratifies the communes, by comparing the commune population density to the average density of the surrounding NUTS 2 region, into one of three levels reflecting population density (i.e. dense, less dense, and not urban). Then, the NUTS 2 totals are disaggregated using a subjectively chosen initial set of weights, re-aggregating the population to the commune level and comparing it to the known total, this way computing a disagreement indicator, and adjusting the weights to reduce the disagreement. The second method estimates the proportion of each CLC class with residential use, overlaying the LUCAS sample on the CLC map, under the assumption that the residential area per person is approximately homogeneous in each commune. The third method uses regression to model the proportion of residential land as a function of the density of the commune, knowing that this density is proportional to the population density. Finally, the fourth method follows several suggestions from previous studies

(Flowerdew et al. (1991); Dempster et al. (1977)), applying the expectation-maximization (EM) algorithm for estimating disaggregation coefficients, and assuming that the population in a land cover class for a given commune follows a Poisson distribution.

Validation was performed by comparing the results with a reference population density grid, concluding that the density attributed to non-urban classes was generally overestimated, and that the performance difference between methods were moderate. It was also detected that, although the dasymetric map was homogeneous, its quality was not, being poorer in the areas where communes were larger (i.e., when the size of the communes was extremely heterogeneous). Some studies suggested improvements to overcome these issues (Eicher and Brewer, 2001), for instance by limiting the density overestimation in non-urban classes, or through the introduction of new demographic information, although this is usually not available at the commune level.

2.2.3 Leveraging Earth Observation Data Together with Other Sources of Ancillary Information

Following an approach similar to Tobler's pycnophylactic interpolation, Mennis (2003) built a methodology for generating a raster-based representation of population using a dasymetric mapping technique that incorporates satellite-derived urban land-cover data, reducing the analytical and cartographic problems associated with aggregated demographic datasets, specifically the modification of the boundaries and/or the scale of the data aggregation. Remote sensing-derived residency data were used, as well as a gridded three-class method that classifies partitions into high-density urban, low-density urban and nonurban areas.

To generate what the author called the *urbanization* dataset, the land-cover data were overlaid with the 1996 Pennsylvania department of transportation road-network data, and the results were photo-interpreted according to land cover and road density. Then, two techniques were proposed to improve previous methods in dasymetric mapping. The first one uses empirical sampling to determine appropriate percentage values that should be assigned to each class, trying to avoid the subjectivity of the process. The second technique is the use of areal weighting to improve the accuracy of the redistribution of population according to the ancillary data, incorporating the difference in area between different ancillary classes within the redistribution calculation, instead of relying exclusively on the nature of the class.

The author claims that the first technique is still somewhat subjective and does not provide a predictive model of the relationship between ancillary data, although previous studies had also shown the difficulty of deriving such models, due to spatial variation in the nature of land

use, urbanization, or other land-cover-based classifications that relate to population density. Still, the author also claims that the proposed techniques are generalizable to a large variety of settings, and that they can be used with other categorical ancillary data that has a demonstrable spatial relationship with the distribution of population.

In an attempt to attenuate the effect of incompatible zone systems caused by the development of rural areas or the redraw of district boundaries, Reibel and Bufalino (2005) used street network data (i.e., TIGER[3] files from the U.S. Census bureau) to derive weights for the interpolation of population and housing unit counts in Los Angeles County, California. A technique of dasymetric areal-interpolation was tested for the complete set of local areas, despite changes in the boundaries, and the results were compared with traditional areal weighting.

The street and road grids were used as a proxy for approximate population and housing unit density surfaces for census tracts in the county, using the following steps. First, three maps were overlaid: the source-zone boundaries, the target-zone boundaries, and the street layer. Then, street weights were computed for each intersection-zone fragment defined by its pair of source and target zones s and t, preserving the pycnophylactic property:

$$W_{st} = \frac{\sum_{vst=1}^{Vst} L_{vst}}{\sum_{vs=1}^{Vs} L_{vs}} \tag{2.12}$$

In Equation 2.12, W_{st} is the weight for a given intersection-zone fragment, while L_{vst} is the length of each street vector vst in that intersection zone, and L_{vs} is the length of each street vector vs in the source zone that belongs to the intersection zone. Finally, the weights are applied to the original source-zone counts attached to each zone, and the weighted estimates for each intersection zone are summed across each target zone.

The authors concluded that the street-weighting method offers some benefits, reducing the errors in estimation when compared to the commonly applied area-weighting technique that assumes a uniform density. However, they also concluded that this reduction is more evident in those areas where the lack of population is reflected in the lack of roads, and not in those with a more developed but nonresidential transportation infrastructure. Also, the use of the street and road grids involves the assumption that the residential population density gradient, at a given distance from the nearest street or road, is constant. The authors have not attempted to use any other attribute information, such as traffic capacity or information about the densities of structures within a given distance from the streets.

[3]http://www.census.gov/geo/maps-data/data/tiger.html

Malone et al. (2012) presented an algorithm, called dissever, that corresponds to a generalised method for downscaling coarsely resolved Earth resource information using available finely gridded data, under the assumption that there is often a nonlinear relationship between the target variable being downscaled and the available covariates. Instead of assuming a linear function to make the correlation, splines and generalized additive modeling were used to model the relations between the indicator that we wish to represent at a fine resolution, and for which we have data at a coarse resolution, and data for other variables, available at a fine resolution. More details about splines and additive models are available in the books by Egerstedt and Martin (2009) and by Hastie and Tibshirani (1990).

In an initialization phase of the dissever algorithm, a resampling is performed, through a nearest neighbor resampling approach, from the coarse grid to the fine grid. The nonparametric smoothing splines f_j that relate the target variable $\hat{t}_m$ to the covariates x_j (i.e., the weights attributed to the covariates) have knots at each of the unique values of $x_{mj},\ m = 1, ..., D$, and are computed through an iterative backfitting algorithm that minimizes the value returned by the penalised residual sum of squares (PRSS) criterion, defined as follows:

$$PRSS = \sum_{m=1}^{D} W_k \left\{ \hat{t}_m - \alpha - \sum_{j=1}^{p} \mathrm{f}_j(x_{mj}) \right\}^2 + \sum_{j=1}^{p} \lambda_j \int \{\mathrm{f}_j^n(t)\}^2 \, \mathrm{d}t_j. \tag{2.13}$$

In Equation 2.13, D is the number of target cells at the fine scale, and W_k is the weighting vector assigned to each target variable, corresponding to a measure of uncertainty estimated in the predictions at the coarse scaled resolutions. The parameter α is a constant, and λ_j is a tuning parameter which controls the trade-off between the fidelity (or the *goodness of data fitting*) term and a penalty term, which is defined by the function curvatures $\int \{\mathrm{f}_j^n(t)\}^2 \, \mathrm{d}t_j$.

In an iteration phase, adjustments are made to the predictions iteratively, so that the coarse grid map is linearly related to the fine grid predictions. The algorithm attempts to ensure that the target variable value for each coarse grid cell equals the average of all target variable values at the fine scale, in each coarse grid cell. The iterations stop when a maximum number is reached or after reaching threshold of 0.001 over the change in the estimated error rate in three iterations. The authors demonstrated the algorithm by downscaling a soil organic carbon map map with a 1x1 km grid resolution to a 90x90 m grid resolution, using available covariate information derived from a digital elevation model, satellite imagery and airborne gamma-spectrometry datasets. The averaged 90x90 km estimates were then compared to the corresponding 1x1 km grid cell values, producing a low root mean square error.

The dissever method assumes that there is an element of uncertainty in the target variable that is being downscaled, unlike other methods that often assume no associated uncertainty. To handle these uncertainties in the empirical process and incorporate them into the downscaling algorithm, higher weights are given to more accurate information. If the uncertainties are not known, downscaling will proceed using equal weights. Additionally, the algorithm can be simply modified to accommodate a user-defined function for capturing relations between variables, for example replacing the nonlinear regression model with other deterministic functions such as linear models, deep neural networks or ensembles of regression trees. The experiments reported in this dissertation have, in fact, used an adapted version of the dissever procedure, specifically considering spatial disaggregation instead of the downscaling of non-additive variables.

Taking advantage of other types of ancillary information, Sridharan and Qiu (2013) developed an areal interpolation model for population data using light detection and ranging-derived building volumes, obtained from LiDAR data, as an ancillary variable. Instead of using two-dimensional areas as the ancillary data, disaggregating population data from the source units (e.g. census block groups) into control units (e.g. residential buildings) and then subsequently reaggregating these data from the control units to the target units (e.g. census blocks), their approach is able to capture the vertical and horizontal distribution of population.

The model is built upon improvements to pixel-based approaches, using a least-squares approximation to the EM algorithm. Innovative methods were proposed for model initialization, for the iterative adjustment (e.g., in order to overcome the spatial heterogeneity problem encountered in earlier approaches) and for the EM stopping criterion, to deal with control units of unequal size. The initialization method is equivalent to applying a binary dasymetric areal interpolation, except that the ancillary variable is building volume rather than area, assigning a source unit population to its buildings in proportion to building sizes (i.e., volumes). The iterative adjustment is carried out in a two-step procedure, estimating the population of the buildings with linear regression (i.e., using their volumes as covariates), and ensuring that the sum of the building-level population estimates equals the census populations of their respective source units (i.e., the pycnophylactic property is maintained), with the following adjustment:

$$\hat{P}_{adj} = \hat{P}_{si} + \left(\frac{(P_s - \hat{P}_s)}{v_s} \right) \times vo_{si} \tag{2.14}$$

In Equation 2.14, $\hat{P}_{si}$ is the population estimate for building i in the source zone s, P_s is the true population of a source zone, $\hat{P}_s$ is the estimated population of a source zone, v is the volume of all buildings in unit s, and vo_{si} is the volume of the that building. The stopping criterion

is then computed with the root-mean-square deviation of the pycnophylactic property between two iterations. The proposed technique was also specifically designed to address the problem of spatial autocorrelation, i.e. the tendency of values in one spatial unit to correlate with those of its neighbors, and nonstationarity issues (i.e., the tendency of the relationship between two variables to vary over space).

As mentioned above, the use of building volumes supports disaggregating not only the horizontal but also the vertical population distribution. Since the model is derived from and applied to a single set of control units, it minimizes the modifiable areal unit problem, in which the relationship between population and ancillary variables quantified at the source units is applied to a different set of spatial units. In addition, residential buildings used in the study are far closer to the spatial resolution of human settlements than residential land use and road networks. However, the authors indicated that the approach needed to be compared with geostatistical models for addressing heterogeneity issues, as well as to evaluate the impact of data inaccuracies such as missing buildings or mislabeling of land use types. To address the variation in the relationship between population and building volumes among different classes, the authors suggested dividing the residence type into single, duplex and multifamily units, this way developing a multiclass spatially disaggregated model.

Lin and Cromley (2015b) also proposed a new polycategorical method that integrates positive aspects of previous regression interpolators that uses locally-varying model parameter estimates, like a geographical weighted regression (GWR)-based (Lin et al., 2011) or a quantile regression (QR)-based (Cromley et al., 2011) approach, for solving areal interpolation problems. Two types of neighborhood heuristics for selecting observations were used, in order to estimate ancillary control densities: one that is spatial (i.e., based on geographic proximity, in which the source zones nearest in distance to the center source zone are selected), and one that is statistical (i.e., based on statistical proximity, in which the source zones closest in their overall population density to the center source zone are selected). In particular, the intuition for using statistical neighborhoods is that control categories in source zones, having similar population densities, should also have similar relationships between ancillary categories and population density.

In the study from Lin and Cromley (2015b), the census tracts and census block groups of nine towns in Hartford County, Connecticut, were used as source zones and target zones, respectively. The pre-classified land cover dataset used as ancillary data was the National Land Cover Database 2006, classified based on four categories - developed, low intensity; developed, medium intensity; developed, high intensity; and deciduous, evergreen, and mixed forest. A second ancillary dataset was compiled based on road networks from the U.S. Census Bureau's

2010 dataset. Then, the density coefficients for the ancillary classes associated with the center source zone were estimated by the following linear program, that tries to minimize the sum of the deviational variables so that the best fit of the density coefficients can be found:

$$\begin{aligned} \text{Minimize} \quad & M\lambda_c^- + M\lambda_c^+ + \sum_i (\lambda_i^- + \lambda_i^+) \\ \text{Subject to} \quad & \sum_j^n \beta_j X_{cj} + \lambda_c^- + \lambda_c^+ = P_c \\ & \sum_j^n \beta_j X_{ij} + \lambda_i^- + \lambda_i^+ = P_i, \text{ for all } i \in N_c \\ & \beta_j, \lambda_c^-, \lambda_c^+, \lambda_i^-, \lambda_i^+ \geq 0 \end{aligned} \tag{2.15}$$

In Equation 2.15, M is a very large positive number, ensuring that the population prediction for the center source zone is exact, while λ_c^- is a deviational variable representing the amount of underestimation for the center source zone, and λ_c^+ is a deviational variable representing the amount of overestimation for the center observation. The parameter β_j is the estimated density value for the j-th ancillary class, X_{cj} is the raster count of j-th ancillary class in the center source zone or the area of that ancillary class, P_c is the population count for the center source zone, and X_{ij} is the raster count of j-th ancillary class in the i-th neighboring source zone or the area of that ancillary class. The parameter λ_i^- is the deviational variable representing the amount of underestimation for the i-th observation by the linear equation, while λ_i^+ is the deviational variable representing the amount of overestimation for the i-th neighboring observation. The parameter P_i is the population count for the i-th neighboring observation, while n is the number of ancillary classes, and N_c is the set of source zones in the center source zone's neighborhood.

Lin and Cromley (2015b) indicated the possibility of developing more complex relationships between population and land cover. Specifically, they argue that one should focus on areas with more complicated patterns of density distribution, to verify if the findings of small margins could be reversed in another situation. The authors referred also that it would be interesting to analyse the tradeoff of model complexity and performance, determining if small marginal improvements could be justified, given the extra computation that is often involved.

2.2.4 Leveraging Mobile Phone Call Records and Other Types of Point-Based Ancillary Information

Instead of using land use or road network information as ancillary data, Zhang and Qiu (2011) introduced a point-based approach to the areal interpolation problem, by using points

locationally associated with the variable of interest, widely available from various databases, as ancillary data. For example, schools, supermarkets, and business centers are often close to population centers, although toxic release inventory sites, landfills, and hazardous material processing plants are often situated away from neighborhoods. A novel method was created to model the connection between the zone variables and the control point locations, leveraging either a linear or a nonlinear exponential function.

The proposed algorithm employs a simple data processing operation, specifically a straight-line distance analysis, to estimate the density distribution. It was assumed that the control points are locationally related to the concentration of a variable of interest, and then its density could be estimated by simple linear or nonlinear exponential functions. The weight of a cell i within each source zone s is calculated based on an inverse distance weight:

$$W_{si} = \left(1 - \frac{\lambda_{si}}{\lambda_{smax}}\right)^{q} \tag{2.16}$$

In Equation 2.16, λ_{si} is the distance from the cell to the closest control point, which is assumed to have the largest influence. The parameter λ_{smax} is the maximum value of λ_{si} within the source zone, and q is a power parameter that controls the degree of local influence (i.e., a higher q suggests that the rate of change in density is higher near a cell). Finally, the calculated density can be smoothed, if necessary. The smoothed density value is computed as the average of the original density values for all the neighboring cells in the smooth moving window.

When analysing the results, the authors concluded that the point-based approach had a much better accuracy than simple approaches that do not use ancillary data. It also produced comparable results with those using land use or road network information, but with tremendous savings in computational cost. In conclusion, point-based methods can also support areal interpolation problems, when the choices are limited by various factors.

In a more recent study, Deville et al. (2014) took advantage of the high penetration rate of mobile phones across the globe, and analyzed the spatiotemporal distribution of phone calls, geolocated to the tower level, this way overcoming many limitations of census-based approaches. In fact, the authors showed how spatial and temporary estimations of population densities can be produced at national scales, and how maps of human population changes can be produced over multiple timescales. These achievements were accomplished by identifying the geographic coordinates of each communication transmission tower, and knowing that the higher the density of the towers, the higher the precision of the phone communication geolocalization. The ability

of this approach to accurately downscale census population data was then compared with a remote sensing method based on the usage of overhead images from distant scenes (e.g., satellite imagery) to derive geographic information, using precision and other quality estimators like the Pearson product-moment correlation coefficient and the root-mean-squared error, with the baseline census-derived population densities as a reference.

The phone call dataset used in this study was obtained from major carriers in Portugal and France, during the following periods: July to August 2007 and November 2007 to June 2008 for Portugal, and May to October 2007 for France. The method described in the paper only considered phone calls, excluding text messages. For each call, the originating and receiving towers, and the day the call was made, were obtained. The estimation of the population density of an administrative unit c_i, based on the phone user density σ_{v_j}, is then a two-step method, which starts by computing the nighttime user density σ_{c_i} with the following equation:

$$\sigma_{c_i} = \frac{1}{A_{c_i}} \sum_{v_j} \sigma_{v_j} A_{(c_i v_j)'} \tag{2.17}$$

In Equation 2.17, A_{c_i} is the area of administrative unit c_i, and $A_{(c_i v_j)'}$ is the intersection area of c_i and the polygon v_j associated with tower j, corresponding to the coverage area of the tower. In the second step of the method, the phone nighttime density values were compared with baseline census-derived population densities available in a training set. Temporal dynamics were derived from phone data using the timestamp associated with each phone call, by dividing the data into calls performed during the day and the night, as well as performed during weekdays or weekends. Seasonal dynamics were analyzed by dividing the calls into a holiday period (July and August) and a working period (all other months).

When comparing the proposed approach against a method leveraging only remote sensing data, the authors found, at the national scale, similar patterns that match baseline data, with major cities being clearly identifiable. However, the proposed method presented a higher spatial detail in a close-up on the capital city of Lisbon, relying on the density of the towers (i.e., substantially higher in urban areas), whereas the remote sensing method depends on the spatial resolution of the geospatial datasets, which often do not capture intraurban variations. Despite concluding that phone use behaviors were relatively stable across space and time in Portugal, it was suggested that the variations in these behaviors could be a result of economic, social, demographic or cultural characteristics. It was also concluded that the application of the method to countries with low penetration rates excludes an important fraction of the population, requiring sensitivity analysis of the impact of phone use inequalities.

In a similar study, Douglass et al. (2015) also argued that telecommunication records provide a reliable estimate of the population density, and showed their correlation to population in a given area. The authors used this information, in conjunction with census data and satellite images, to create high-resolution population estimates for Milan, Italy, in time and space. The first step of the study investigated the proportionality of calling activity to population and how it varied over different geographic scales. Then, a model for predicting the population of a given area, and the percentage of that population which is of foreign nationality, was proposed.

To disaggregate the population measures obtained from the Italian national institute of statistics, the authors first developed a land use map of the region, classifying each 15x15 m square into classes corresponding to buildings, vegetation, water, road/pavement, and railroads. The authors have specifically trained a random forest classifier (Breiman, 2001), using OpenStreetMap data as labeled training examples. The measures of telecommunication activity were provided by Telecom Italia as part of a big data challenge. The data was acquired in Milan between November 1, 2013 and January 1, 2014, and it was also divided into calls out, calls in, text messages sent, text messages received, and Internet activity. Then, to compute the correlation between calling activity and population, an elementary model was proposed, in which the call volumes are assumed to be directly proportional to the population estimates. Since each communication type correlates most strongly with population during the hour from 10 am to 11 am, the data corresponding to that period was used as predictions in the model. Finally, with the goal of producing up-to-date population data in between census years, the authors proposed a method that used random forest regression (Breiman, 2001).

In conclusion, Douglass et al. (2015) referred that a similar analysis could be made to create sub-population estimates by age, gender, and ethnicity. However, the authors have also shown that mobile telecommunication activities are most useful in conjunction with other spatial measures of human activity, tuned to local conditions (i.e., the model parameters developed for one region or country cannot be applied blindly to another region).

2.2.5 Leveraging Geo-referenced Social Media as Ancillary Information

Longley and Adnan (2015) developed a classification framework that used geotagged social media data (i.e., data that documents important aspects of the daily activities of millions of users, as well as social attitudes and opinions) as indicators of spatial behaviour. Specifically, highly disaggregated information collected from Twitter helped to develop geo-temporal analysis methods to characterize the area of Greater London in terms of flows of people, combining models

of individual characteristics gathered from social media with conventional measures of land use morphology and nighttime residence. The separate study from Steiger et al. (2015) also examined the possibility of using Twitter data, presenting a spatio-temporal and a semantic analysis framework for geo-referenced tweets from the same area of Greater London. Particularly, two research questions were formulated: (1) what are the work-related and home-related activities that reflect typical collective human behavior; and (2) to what degree can the observed tweet clusters be regarded as a proxy of human social activities and correlate with available residential and workplace populations from official UK census data.

Also using social media information, Jiang et al. (2015) demonstrated a process to collect, unify, classify and validate online point-of-interest (POI) data, and developed methods to use these data to estimate disaggregated land use at a very high spatial resolution (i.e., at the census block level), using part of the Boston metropolitan area as an example. The point-of-interest information collected from social networks (i.e., specific point locations that a considerable group of people find useful or interesting) is suggested to, along with modern techniques for geo-processing, derive disaggregated data that represent activity patterns in cities.

The data sources used in the study by Jiang et al. (2015) included employment by category at the aggregate census block group level, volunteered POI information, and geographic boundaries of both the aggregate and disaggregated area of analysis. Then, the authors proposed to use Web mining and machine learning techniques to automatically collect POIs from different sources, and to classify them into a standard taxonomy, which was essential since these POIs were collected from different sources. A POI Matching algorithm (Cohen et al., 2003) was used to map POIs from Yahoo![4] to those from Dun & Bradstreet (D&B)[5], and to generate training data for the machine learning algorithms, by identifying similar names, ignoring misspelling errors and abbreviations, and by comparing POIs according to their names, websites and geospatial distances. Weka (Witten and Frank, 2005), a data mining platform that provides a portfolio of classification algorithms, was used to classify POIs into the different North American Industry Classification System (NAICS) levels, specifically the two-digit codes that allow analysis of economic sectors, and the six-digit codes that specify categories of business establishments. Finally, a local *maximum likelihood estimation* (MLE) method was applied to disaggregate the block group level aggregates to block level land use estimations, based on the assumption that the estimates of the employment sizes within different blocks are proportional to the number of POIs within the blocks (i.e., the employment sizes of a certain category within a block group are

[4]http://www.yahoo.com
[5]http://www.dnb.com

independent and identically distributed). In other words, the share of the estimated employment size of a block in a block group is equal to the share of POIs of the block in the block group.

The authors indicated that, since POI information was extracted from online platforms, the coverage and accuracy of the information depends on the completeness of online public sources and on the consistency of public categories. Other issues relate to the fact that the frequency of obtaining updated POI data varies, as well as to the restrictions on the massive usage of Volunteered Geographic Information, according to the legal terms of the different data providers. However, the combination of different online sources can help in the reduction of the gaps between the total business establishments and the available information online. In general, for cities without resources to update business establishment data, the proposed method provides an alternative for developing accurately and timely disaggregated land use estimations, as well as for analyzing urban and regional economies.

Also taking advantage of the open access to geographic data generated within social networks, Lin and Cromley (2015a) used social media information as a control layer for areal interpolation of population. Their study evaluated the effectiveness of geo-located nighttime Twitter data as a single source of ancillary information, or as an integrated source to be combined with other control variables, such as information on land coverage, road network, and residential parcels. The study area was a nine-town region within Hartford County, Connecticut, characterized by different types of land use with various population densities (e.g., medium urban fabric, dense urban center, and forest areas).

In their study, the nested hierarchy of the census tract and block groups were used as source and target zones, respectively. The population datasets were obtained from the U.S. CENSUS Bureau's 2010 dataset, while the land coverage data were classified based on the Landsat Thematic Mapper satellite imagery. The road networks were acquired from the U.S. Census Bureau's 2010 TIGER/Line dataset, and the parcel data was obtained from the Connecticut Capitol Region Council of Governments. The geo-located tweets were captured in near real-time using the streamR[6] package for the R environment, and captured within two time intervals, in order to identify as many residential places as possible and to mitigate the bias of tweets sent during each day. The authors first selected the geo-located tweets sent between 6 PM and 8 AM, and then kept only one tweet from the active users that sent more than one tweet in nearby locations (i.e., within a three hundred meters radius). The total population was then estimated for the Twitter data, as well as for the other datasets. For the Twitter dataset, two different

[6]http://cran.r-project.org/web/packages/streamR

weighting surfaces were created, namely one using a linear distance decay function, and another using a nonlinear model. The estimated density value $\hat{D}_{si}$ for cell i in the source zone s was computed using the following equation:

$$\hat{D}_{si} = \frac{W_{si}}{\sum_{i=1}^{N_s} W_{si}} \times P_s \tag{2.18}$$

In Equation 2.18, N_s is the cell count in source zone s, while W_{si} is the weighting value for cell i in the source zone s, and P_s is the population count in source zone s. The estimated population in each target block group was obtained by aggregating all the cell values in that zone. For the land coverage, road network and parcel datasets, the developed areas, road buffer areas and residential parcels were respectively integrated as the control variables in a typical binary dasymetric model, to obtain the corresponding estimates at the block group. Finally, age-specific populations were also estimated to determine if the control data was more spatially correlated with the distribution of certain age groups of the population.

The authors concluded that geo-located tweets do not perform as well as single control data, when compared to the three other control layers, for total population and for all age-specific population groups. However, Twitter data were found to have an enhancing effect on other control data, specifically for age groups with a high percentage of Twitter users, increasing the variation in the densities within a residential area. Lin and Cromley (2015a) also concluded that the combination of more control variables does not necessarily lead to an improved areal interpolation performance. In fact, some control data may weaken others (e.g., the addition of geo-located tweets for the interpolation of the elderly population can lead to a loss of accuracy).

2.2.6 Overview on the Related Work

Table 1 presents an overview of the previous studies that were surveyed in this section, summarizing their conclusions and illustrating similarities between them.

According to Wu et al. (2005), the methods reported for population downscaling can be grouped into two categories: areal interpolation and statistical modeling. While areal interpolation methods are designed to transfer data from one set of spatial units to another, statistical modeling has the primary objective of creating relationships between population and other variables, trying to make easier the task of estimating the total population for an area. Areal interpolation often uses census population data to obtain a population surface, unlike statistical modeling, that uses socioeconomic and/or Earth observation variables to apply theoretical

Paper	Techniques	Data
Goodchild et al. (1993)	Mass-preserving areal interpolation	Socioeconomic and hydrologic
Elvidge et al. (1997)	Dasymetric mapping	Visible-near infrared emissions
Doll et al. (2000)	Dasymetric mapping	Nighttime lights and statistical information on human environment
Mennis (2003)	Dasymetric mapping	Land cover and road network
Reibel and Bufalino (2005)	Dasymetric mapping	Road network
Briggs et al. (2007)	Dasymetric mapping	Land cover and nighttime lights
Gallego (2010)	Dasymetric mapping and regression	Land cover
Zhang and Qiu (2011)	Dasymetric mapping	Zero-dimensional points
Malone et al. (2012)	Splines and generalized additive modeling	Covariates from digital elevation model, satellite imagery and airborne gamma-spectrometry
Sridharan and Qiu (2013)	Dasymetric mapping and regression	LiDAR building volumes
Deville et al. (2014)	Dasymetric mapping	Telecommunication records
Lin and Cromley (2015b)	Dasymetric mapping and regression	Land cover and road networks
Douglass et al. (2015)	Dasymetric mapping	Telecommunication records and land cover
Jiang et al. (2015)	Dasymetric mapping	Census, points-of-interest and geographic boundaries
Lin and Cromley (2015a)	Dasymetric mapping	Census, tweets, land coverage, road network and residential parcels

Table 2.1: The spatial downscaling/disaggregation approaches that have been surveyed.

methods in urban geography for population estimation.

Areal interpolation methods are sometimes separated into simple or intelligent ones (Lin and Cromley, 2015b). While simple areal interpolation methods transfer data from source zones to target zones without using ancillary data, intelligent methods use ancillary data that provide correlated density surfaces, this way improving estimation accuracy. Two types of ancillary datasets that are commonly used are remotely sensed land cover data (Mennis (2003) and Douglass et al. (2015)) and road network data (Mennis (2003) and Reibel and Bufalino (2005)). In addition to these two, nighttime light emissions (Doll et al., 2000) and satellite observed visible to near infrared emissions (Elvidge et al., 1997) can also be used as an indicator for population size. In more recent studies, Zhang and Qiu (2011) and Jiang et al. (2015) used point data for businesses and other landmarks, while Deville et al. (2014) and Douglass et al. (2015) used geo-referenced telecommunication records, with the same purpose. The dasymetric method is commonly regarded as a more accurate approach, in comparison to simple areal weighting, to guarantee that the used ancillary information gives a meaningful help in the context of spatial disaggregation.

Statistical modeling can sometimes be incorporated into dasymetric methods, because it can also be used to disaggregate several types of indicators. In fact, the main challenge in dasymetric disaggregation is to devise an appropriate set of weights that can be applied to the land parcels, and regression analysis is commonly used to define these weights. Malone et al. (2012) proposed the dissever algorithm, that uses splines and generalized additive modeling to fit a non-linear relationship between a target variable and predictive covariates, while Briggs et al. (2007), Gallego (2010), Sridharan and Qiu (2013), and Lin and Cromley (2015b) used simpler forms of regression analysis with the same purpose.

The novel spatial disaggregation method that is described in this dissertation builds on and extends the previous research described in this chapter, particularly the downscaling method advanced by Malone et al. (2012). Unlike most aforementioned studies, the focus of the experiments reported in this dissertation was not so much in the disaggregation of population densities, but instead in the computation of high spatial resolution estimates for other socio-demographic indicators. Since some of the experiments used social media data in a dasymetric mapping technique, the work has some similarities with the study conducted by Lin and Cromley (2015a). However, instead of using nighttime tweets, the ancillary data used in these experiments are acquired from another social media source (i.e., Flickr), and not restricted to a specific time of the day (i.e., Flickr photos were not filtered according to particular time periods).

3 Spatial Disaggregation Based on Regression Analysis

This chapter details the considered spatial disaggregation approach. First, Section 3.1 outlines all the steps of the procedure. Then, the regression algorithms used to combine multiple sources of ancillary data, such as standard linear models, generalized additive models, the cubist method, robust linear regression, and geographically weighted regression, are presented in Section 3.2. The ancillary sources of information are, in turn, described in Section 3.3. Finally, Section 3.4 overviews the contents of this chapter.

3.1 The Proposed Hybrid Algorithm

Both the dasymetric mapping and the pycnophylactic interpolation methods have solid theoretical foundations, as well as strong empirical supports in population-estimation research. Each of these methods has its own strengths, but also suffers obvious shortcomings. For instance, pycnophylactic interpolation warrants a smooth surface in the study area, without any presumption of uniform distribution. However, the method does not draw on any ancillary information about the real spatial distribution, so that its estimation accuracy cannot benefit from useful information that may be easily available.

In the context of this work, I developed an hybrid approach that takes advantage of the strengths and that remedies the flaws of both methods, following the general ideas presented by Kim and Yao (2010) and by Malone et al. (2012). Similarly to Kim and Yao (2010), I propose to combine pycnophylactic interpolation with dasymetric mapping (i.e., binary dasymetric disaggregation in the case of Kim and Yao (2010), whereas I use general dasymetric mapping with basis on a regression approach, taking inspiration on the work outlined by Malone et al. (2012)). The hybrid approach consists of three main consecutive logical steps, namely a simple iterative pycnophylactic interpolation process for generating a preliminary data redistribution, followed by a dasymetric mapping (i.e., proportional and weighted areal interpolation, using population as ancillary data), and finally using a disseveration-based procedure to combine both previous estimates with other ancillary variables. The general procedure is detailed next, trough an enumeration of all the individual steps that are involved in data disaggregation:

1. Produce a thematic map for the variable to be disaggregated by associating the quantities, linked to the source regions, to geometric polygons representing the corresponding regions;

2. Create a raster representation for the study region, with basis on the thematic map from the previous step and considering a resolution r. This raster, referred to as T^p, will contain smooth values resulting from a pycnophylactic interpolation procedure (Tobler, 1979). This interpolation algorithm starts by assigning cells to the corresponding values in the original thematic map, using a simple mass-preserving areal weighting procedure (i.e., it redistributes the aggregated data with basis on the proportion of each source zone that overlaps with the target zone). Interactively, each cell's value is replaced with the average of its neighbours in the target raster. The values of all cells are finally adjusted within each zone proportionally, so that each zone's total in the target raster is the same as the original total (e.g., if the total is 10% lower than the original value, the value of each cell is increased by a factor of 10%). The procedure is repeated, until no more significant changes occur. The resulting raster with a resolution r corresponds to an initial estimate for the disaggregated values;

3. Overlay four rasters P^1, P^2, P^3 and P^4, also using the same resolution r, on the study region from the original thematic map and from the raster produced in the previous step, respectively with information regarding (i) population counts, (ii) nighttime light emissions, (iii) land coverage classification, and (iv) OpenStreetMap road network density. These rasters will be used as ancillary information for the spatial disaggregation procedure. Prior to overlaying the data, the four different raster data sources are normalized to the resolution r, through a simple interpolation procedure based on taking the mean of the different values per cell (i.e., in the cases where original raster had a higher resolution), or the value from the nearest/encompassing cell (i.e., in the cases where the original raster had a lower resolution);

4. Overlay two other rasters P^5 and P^6 over the study region, again using the same resolution r and with ancillary information derived from the rasters in the previous step. Specifically, these two new rasters encode (i) the distance from a given cell to the nearest cell with a land coverage type equal to water, and (ii) the distance from a given cell to the nearest cell containing a road or a street segment. Raster P^5 is thus derived from raster P^3 with land coverage information, whereas raster P^6 is derived from raster P^4 with OpenStreetMap road network density. These two rasters will also be used as ancillary information for spatial disaggregation;

5. Overlay another raster T^d over the study region, with the same resolution r used in the rasters from the previous steps. This raster will be used to store the estimates produced by a simple spatial disaggregation procedure based on dasymetric mapping (i.e., a method based on proportional and weighted areal interpolation). For producing these estimates, the total value is weighted, for each source zone in the original thematic map, according to the proportion between the population values available for the corresponding cell in raster P^1, and the sum of all the values for the given source zone in the same raster. This is essentially a proportional and weighted areal interpolation method, corresponding to the following equation, where T_t^d is the estimated count in target zone t, where S_s is the count in source zone s, where P_t is the population count in target zone t, and where P_s is the population count in source zone s:

$$T_t^d = \sum_{s=1}^{S} \left(\frac{P_t}{P_s} \times S_s \right) \tag{3.1}$$

6. Create yet another raster P^7 with the distribution of social media usage (i.e., the number of social media contents) under the geographic region, using also the resolution r. After assigning to each cell the sum of all the geo-referenced items that fall into that area, a kernel density estimation procedure is applied to smooth the values and obtain a density map to be used as ancillary data;

7. Collect a sample of cells in the fine resolution grid, in order to latter fit regression models. The sampling procedure can, for instance, be performed using the R function spsample[1], that supports regular (i.e., systematically aligned) sampling, which can evenly represent the entire geographic region while at the same time avoiding the problem of spatial auto-correlation, as well as clustered sampling (i.e., the same number of samples are collected from groups of points assumed to have different characteristics). Two of the experiments reported in this dissertation have specifically used regular sampling, whereas the others considered the entire set of available points as training data;

8. Create a final raster overlay, through the application of an intelligent dasymetric disaggregation procedure based on disseveration, as proposed by Malone et al. (2012), and leveraging the rasters from the previous steps. Specifically, the thematic map from Step 1 is considered as the source data to be disaggregated, while raster T^p from Step 2 is considered as an initial estimate for the disaggregated values. Rasters P^1, P^2, P^3, P^4,

[1] `http://www.rdocumentation.org/packages/sp/versions/1.2-3/topics/spsample`

P^5, P^6, P^7 and T^d are seen as predictive covariates. The regression algorithm used in the disseveration procedure is fit using the data sample from the previous step, and applied to produce new values for raster T_p;

9. The values returned by the downscaling method from Malone et al. (2012) are proportionally adjusted for all cells within each source zone, so that each source zone's total in the target raster is the same as the total in the original thematic map (e.g., again, if the total is 10% lower than the original value, increase the value of each cell by a factor of 10%);

10. Steps 7 to 9 are repeated, iteratively executing the disseveration procedure that relies on regression analysis to adjust the initial estimates T^p from Step 2, until the estimated values converge or until reaching a maximum number of iterations.

Notice that some of the experiments reported in this dissertation have not used all the predictive covariates that were listed in the previous enumeration (e.g., many of the experiments did not use the Flickr data). In most of the experiments, the sampling procedure of Step 7 was also ignored (i.e., the regression model was fit leveraging all available data).

The above procedure was implemented through the programming language of the R[2] project for statistical computing, given that there are already many extension packages[3] for R concerned with the analysis of spatial data, facilitating the usage of geospatial datasets encoded using either the geometric or the raster data models (Bivand et al., 2012). I specifically relied on the R packages named pycno[4] and dissever[5], which respectively implement the pycnophylactic interpolation algorithm from Tobler (1979) used in Step 2, and the downscaling procedure based on regression analysis and disseveration, that was outlined by Malone et al. (2012) and that is used in Step 8. By extending dissever, I could easily perform disaggregation experiments with different types of regression models, such as ensembles of decision trees as used by Stevens et al. (2015), or generalized additive models are originally used by Malone et al. (2012). The latest version of dissever was already internally using the caret[6] package, in terms of the implementation of the regression models. The caret package (Kuhn, 2008), short for classification and regression training, contains numerous tools for developing different types of predictive models, facilitating the realization of experiments with different types of regression approaches in order to discover the relations between the target variable to disaggregate, and the available covari-

[2] http://www.r-project.org
[3] http://cran.r-project.org/web/views/Spatial.html
[4] http://cran.r-project.org/web/packages/pycno/index.html
[5] http://github.com/pierreroudier/dissever
[6] http://cran.r-project.org/web/packages/caret

ates. The experiments reported in this dissertation have used standard linear regression models, generalized additive models (Hastie and Tibshirani, 1990), an approach based on ensembles of decision trees that is typically referred to as cubist (Quinlan, 1992), robust linear regression models (Huber et al., 1981), and geographically weighted regression models (Lin et al., 2011). Geographically weighted regression is not included in the caret package, but extending dissever in order to use this method was relatively straightforward. The updated implementation of dissever, with all the referenced adaptations, is now publicly available[7].

3.2 The Considered Regression Algorithms

As mentioned in the previous section, this dissertation reports on experiments with different types of regression models, within the disseveration procedure.

In standard linear regression, a linear least-squares fit is computed for a set of predictor variables (i.e., the covariates) in relation to a dependent variable (i.e., the disaggregated values). The well known linear regression equation corresponds to a weighted linear combination of the predictive covariates, added to a bias term, and tries to estimate the set of regression parameters by minimising the sum of the residual squares (i.e., the differences between the estimates and the observations). A single equation is computed, with global parameters that reflect a relation assumed to be stationary over space, i.e., the parameters are applied equally over the whole region. Standard linear regression was already described in Section 2.1.

In generalized additive models, the dependent variable values are also predicted from a linear combination of predictor variables, but these are instead connected to the dependent variable via a link function, which nonetheless may simply correspond to the identify function. Instead of a single coefficient for each variable in the model (i.e., for each additive term in the linear combination), a function (e.g., a cubic spline smoother) is estimated for each predictor, to achieve the best prediction of the dependent variable values. Instead of estimating single parameters, in generalized additive models we have that a more general function is found, that relates the predicted values to the predictors, effectively allowing for some degree of non-linearity. Detailed descriptions of how generalized additive models are fitted to data can be found in the book from Hastie and Tibshirani (1990).

The cubist approach, originally proposed by Quinlan (1992), is instead based on combining decision trees with linear regression models, again allowing for some degree of non-linearity. The

[7] `http://github.com/bgmartins/dissever`

leaf nodes in these decision trees contain linear regression models based on the predictors used in previous splits. There are also intermediate linear models at each step of a tree, so that the predictions made by the linear regression model, at the terminal node, are also smoothed by taking into account the predictions from the linear models in the previous nodes, recursively up the tree. The tree-based cubist approach is also normally used within an ensemble classification scheme based on committees, in which a series of trees is trained sequentially with adjusted weights. The final predictions result from the average of the predictions from all committee members in the ensemble. Some of the major innovations of the cubist model result from the application of post-processing techniques, after the initial tree has been grown, like (i) a simplification of the linear models, by using a greedy search to remove variables that contribute little to the model, so that the estimated error is minimised, (ii) an examination of each of the non-leaf nodes of the model, selecting the simplified linear model or the model subtree, depending on which has the lower estimated error, and (iii) a smoothing process on each leaf, to reflect the predicted values at nodes along the path from the root to that leaf. To tune the cubist model over the number of committees and neighbors (i.e., how many, if any, neighbors should be used to correct the model predictions), the training function in the caret package implements cross-validation procedures, allowing one to find appropriate settings for these parameters.

Robust regression approaches are designed to circumvent some of the limitations that traditional methods have. In fact, ordinary least squares regression is built on particular assumptions over the data. If those properties are not true, misleading results can occur. For example, the presence of outliers (i.e., observations that do not follow the pattern of the other data) or heteroscedasticity (i.e., the variance of the error is not constant for all the data) can strongly impact the results of ordinary least-squares regression. One solution is to remove influential observations from the least-squares fit. Another approach, termed robust regression, is to employ a fitting criterion that is not as vulnerable to unusual data. Two popular alternatives of robust regression correspond to (i) least squares approaches that use absolute deviations rather than square ones (Rousseeuw and Leroy, 1987), and (ii) parametric approaches that assume that the residuals follow literal heavy-tailed distributions, like t-distributions, or a mixture of normal ones (Lange et al., 1989). The experiments reported in this dissertation used the most common iterative approach from the first category, named re-weighted least squares, with an M (i.e., maximum likelihood type)-estimator. The M-estimation procedure was first introduced by Huber (1973), and is a technique that tries to minimize the sum of each residual contribution, that is given by its value multiplied by an assigned weight. After computing initial estimates $b^{(0)}$ through a simple least-squares estimation, new weighted-least-squares values are calculated at

each iteration based on the previous weight values, as follows:

$$b^{(t)} = \left[X'W^{(t-1)}X\right] X'W^{(t-1)}y \tag{3.2}$$

In the previous equation, X' is the model matrix, that contains each of the instances x_i' as its rows, and $W^{(t-1)}$ is the current weight matrix. The procedure is repeated until the estimated coefficients converge (i.e., the predicted values do not change significantly after one iteration).

Finally, when considering Geographically Weighted Regression (GWR), the focus is on the identification of several local relationships between variables, i.e., locally weighted regression coefficients concerning different geographic zones. The reasons why parameters are expected to be different in distinct parts of a study region concern random sampling variations, or different intrinsically disparities of administrative, political or contextual issues. Instead of estimating a single regression equation, with global parameters that are assumed to apply equally over the whole region, GWR takes into account the phenomena of spatial nonstationarity, introduced by Fotheringham et al. (1998). GWR consists in a weighting scheme that specifies a continuous weighting function, depending on the distance between sampled points and a control location, as well as on a parameter (i.e., a bandwidth) that ensures higher influence of nearer sampled points, instead of those farther away. The selection of an appropriate bandwidth is crucial, since a higher value makes the estimated parameters uniform (i.e., without an appropriate representation of the local variations), and smaller values cause a big dependency on observations near the control location. This way, one possibility is to choose the bandwidth through the minimization of the sum of squared errors, although this approach often produces values that tend to zero once the bandwidth value is so small that all the points except the control location become negligible. Another solution, that is employed in the experiments reported in this dissertation, is to perform a cross-validation on the data as the bandwidth becomes very small, by calibrating the model only with samples near to the control point and not at the point itself. To perform the experiments using GWR, we specifically relied on the the R package GWmodel[8]. Details on the geographically weighted regression procedure were already given in Section 2.1.

I effectively experimented with these different regression models to measure their impact on the disaggregation performance for different types of variables. In cases where the target variable has a smooth and nearly linear dependence on the covariates, a standard linear regression model will probably perform better than more sophisticated non-linear approaches (e.g., an approach

[8] `http://cran.r-project.org/web/packages/GWmodel`

based on a combination of multiple decision trees, which will attempt to approximate the linear curve with an irregular step function). In the presence of multi-collinearity, or for more complex relationships between the target values and the covariates, non-linear models can perhaps offer a better performance. Also, if the relationships between an indicator and the ancillary data can be expected to vary across the study region, a state-of-the-art method like geographically weighted regression can produce significantly better results, since this will calibrate multiple local parameters. Finally, if the data have outliers, robust regression is probably a better solution over simple least squares regression.

3.3 Sources of Ancillary Data

The ancillary information that was used in the disaggregation procedure refers to population counts, nighttime light emissions, land coverage classification, OpenStreetMap road network density, and social media usage.

The data regarding population statistics were obtained from the Gridded Population of the World (GPW[9]), a well-known dataset depicting the distribution of human population across the globe, providing globally consistent and spatially explicit (i.e., disaggregated) human population information. The current version of the dataset was constructed from national or subnational input units (i.e., from low-level administrative units from the different countries) of varying resolutions, through a complex spatial disaggregation procedure. The initial version of GPW, which was released in 1995 and used a simple pycnophylactic spatial disaggregation method for population data (Tobler et al., 1995), resulted from a discussion at the 1994 workshop on global demography, where there was consensus that a consistent global database of population totals, in raster format, would be invaluable for interdisciplinary research. The dataset was then continually revised over the years.

The grid resolution considered for the GPW dataset is of 30 arc-seconds per cell, or 1km at the Equator, although aggregates at coarser resolutions are also provided. Separate grids are available with representations for the population counts and for the density per grid cell. Population data estimates (in 2015, when GPWv4 was released) are provided for 2000, 2005, 2010, 2015 and 2020, extrapolating from data collected in the 2010 round of censuses, which occurred between 2005 and 2014. The experiments reported in this dissertation used the count data projected to the year of 2010, with the resolution of 30 arc-seconds per cell.

[9] `http://beta.sedac.ciesin.columbia.edu/data/collection/gpw-v4`

As for the ancillary information regarding nighttime light emissions, the experiments reported in this dissertation used the publicly available VIIRS Nighttime Lights-2012 dataset[10], maintained by the Earth Observation Group of the NOAA National Geophysical Data Center. Since 1992, the NOAA U.S. National Geophysical Data Center produces and provides a long time-series and global dataset of annual nighttime satellite images from the U.S. Air Force Defense Meteorological Satellite Program (DMSP), using the Operational Linescan System (OLS). In the past, the distribution of artificial light from these images has been used in many different studies, as a proxy for urbanisation, population density, economic activity, and armed conflict, as well as to assess the spatial extent of light pollution itself (Li et al., 2016; Elvidge et al., 2009). I specifically used the global cloud-free composite of VIIRS nighttime lights, which was generated using VIIRS day/night band (DNB) observations collected on nights with zero moonlight, respectively on 18-26 of April 2012, and on 11-23 of October 2012. Cloud screening was done based on the detection of clouds in the VIIRS M15 thermal band, and the product has not been filtered to subtract background noise, or to remove light detections associated with fires, gas flares, volcanoes or aurora. The raster data, available at a resolution of 15 arc-seconds per cell, consists of floating point values calculated by averaging the pixels deemed to be cloud-free.

On what regards land coverage information, the experiments reported here used the standard Corine Land Cover (CLC) data product[11], which is based on satellite images as the primary information source, and whose technical details are presented in the report by Heymann and Bossard (1994). I specifically used data for the year of 2012, made available on a 250×250 meter resolution. The 44 different classes of the 3-level Corine nomenclature that is considered in the original product (e.g., classes for water bodies, artificial surfaces, agricultural areas, etc.) were converted into a real value in the range $[0, 1]$, which encodes how developed is the territory corresponding to a given cell (i.e., cells with the class water bodies were assigned the value of zero, cells corresponding to wetlands were assigned the value of 0.25, different types of forest and semi-natural areas were assigned the value of 0.5, agricultural areas were assigned the value of 0.75, and artificial surfaces were assigned the value of one). This conversion from categorical to numeric values makes it easier to explore land coverage within different types of regression modeling methods (e.g., this procedure is appropriate for standard linear regression models, where categorical variables would otherwise have to be encoded through a more complex procedure, for instance through the use of one different variable for each possible category, with a value of one if the case falls in that category and zero otherwise). Besides the raster encoding

[10] `http://ngdc.noaa.gov/eog/viirs/download_monthly.html`
[11] `http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012/view`

land development, the CLC dataset was also used to produce a second raster with derived information, encoding the distance towards the nearest water body (i.e., the distance towards the nearest CLC cell assigned to the class *water bodies*). This derived dataset may be easier to explore within particular types of regression methods.

On what regards OpenStreetMap information, I used the methodology associated to a study put forward by Martin Raifer in 2015, regarding the most densely mapped regions in OpenStreetMap[12]. From a shapefile containing OpenStreetMap road network information[13], I computed a raster with a resolution of 30 arc-seconds per cell, and where each cell is associated to the total number of nodes from the street network in that area. OpenStreetMap information was also used to produce an additional dataset with derived information, again with a resolution of 30 arc-seconds per cell and encoding the distance from a given position (i.e., from a given cell) towards the nearest road or street segment.

Finally, the social media data that were considered in some of the experiments (i.e., not all the tests involved the application of Step 6 from the algorithm put forward on Section 3.1) were originally made available in the context of the location estimation subtask from the MediaEval Placing Task[14], which concerned with the development of data-driven methods to estimate the latitude/longitude coordinates at which a photograph was taken. A large sample of geo-referenced photos was drawn from Flickr for this joint evaluation. Specifically, I relied on the publicly available information regarding the 2013[15] and the 2015[16] editions of the placing task, that integrates various aspects of multimedia: textual metadata, image content, location, time, users and context. The MediaEval 2013 Placing Task dataset contains nearly nine million images. The image metadata was crawled through the Flickr API, while visual features were extracted with the open-source LIRE library for content based image retrieval (Hauff et al., 2013). On the other hand, the training set from the MediaEval 2015 Placing Task contains 4672382 photos and 22767 videos, providing several visual, aural and motion features to the participants. To compute a raster containing the density of the Flickr photos from both MediaEval datasets, I first used the coordinates of all the images (i.e., from the 2013 and the 2015 editions, although originally published on Flickr at different years) to produce a map with their counts per each raster cell. After that, a kernel density estimation function was applied to the previous raster, using an appropriate bandwidth selector that implements biased cross-validation (i.e., the bandwidth

[12] `http://github.com/tyrasd/osm-node-density`
[13] `http://download.geofabrik.de/europe.html`
[14] `http://www.multimediaeval.org`
[15] `http://www.st.ewi.tudelft.nl/~hauff/placingTask2013Data.html`
[16] `http://repository.tudelft.nl/islandora/object/uuid:ec44e1d0-1228-463a-a6c2-e693b8091bc1`

selector named *bw.bcv* from the R package named *stats*[17]), to smooth the count data.

3.4 Overview

This chapter presented the novel approach for spatial disaggregation that was considered in this dissertation. Concretely, it first presented the general algorithm that allows the combination of multiple sources of ancillary information to aid in the disaggregation procedure. Then, the chapter presented the different regression models that were tested, with particular focus on state-of-the-art techniques that allow the calibration of multiple local regression coefficients under the geographic territory. It finally presented the ancillary data sources used to support the disaggregation of socio-economic indicators, with emphasis on those that were less studied in the literature, like social media data collected from Flickr.

[17] `http://stat.ethz.ch/R-manual/R-devel/library/stats/html/bandwidth.html`

4 Experimental Evaluation

This chapter presents the experimental evaluation of the disaggregation procedure, which involved tests with Portuguese socio-economic indicators, and tests with tourism indicators concerning the territories of Belgium and France. The general evaluation methodology is presented first, together with the description of the evaluation metrics used to measure the error of the spatial disaggregation procedure (Section 4.1). Next, the results for the Portuguese case study are presented and discussed (Section 4.2), followed by the results for the Belgian and French case studies (Section 4.3). The analysis of the results is supported on (i) graphs showing the correlation between the indicators and the ancillary datasets, (ii) tables with disaggregation error values, and (iii) maps with the distribution of the errors over the territory.

4.1 Methodology and Evaluation Metrics

It should be noted that spatial disaggregation is never an error-free process, and errors introduced during disaggregation can be propagated to subsequent spatial data analysis steps. The typical accuracy assessment strategy for spatial downscaling/disaggregation methods involves aggregating the target zone estimates to either the source or some intermediary zones, and then comparing the aggregated estimates against the original counts. The results for the comparison can be summarized by various statistics, such as the Root Mean Square Error (RMSE) between estimated and observed values, or the Mean Absolute Error (MAE). The corresponding formulas for both these metrics are as follows.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{4.1}$$

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n} \tag{4.2}$$

In the previous equations, $\hat{y}_i$ corresponds to a predicted value, y_i corresponds to a true value, and n is the number of predictions. Using multiple error metrics can have advantages, given that

individual measures condense a large number of data into a single value, thus only providing one projection of the model errors that emphasizes a certain aspect of model performance. For instance Willmott and Matsuura (2005) proved that the RMSE is not equivalent to the MAE, and that one cannot easily derive the MAE value from the RMSE (and vice versa). While the MAE gives the same weight to all errors, the RMSE penalizes variance, as it gives errors with larger absolute values more weight than errors with smaller absolute values. When both metrics are calculated, the RMSE is by definition never smaller than the MAE. Chai and Draxler (2014) argued that the MAE is suitable to describe uniformly distributed errors, but because model errors are likely to have a normal distribution rather than a uniform distribution, the RMSE is often a better metric to present than the MAE. Multiple metrics can provide a better picture of error distribution and thus, in this dissertation, the results are presented in terms of the MAE and RMSE metrics. The results are also reported in terms of the Normalized Root Mean Square Error (NRMSE) and the Normalized Mean Absolute Error (NMAE), in which the values of the RMSE and MAE are divided by the amplitude of the true values (i.e., the subtraction of the maximum true value by the minimum true value). This normalization can facilitate the comparison of results across variables.

To get some idea on the errors that are involved in the experiments reported in this dissertation, I experimented with the disaggregation of data originally reported at the level of large territorial divisions to the raster level, latter aggregating the estimates to the level of civil parishes or municipalities (i.e., taking the sum of the values from all raster cells associated to each division), and comparing the aggregated estimates against known values.

It should be also noted that not all of the regression models and/or ancillary variables were used in the entire set of experiments, regarding the different case studies, that are reported in this dissertation.

4.2 Experiments with Portuguese Socio-Economic Indicators

In the Portuguese case study, socio-economic data pertaining to the Portuguese territory and its administrative units were used, mostly at the level of civil parishes. The data are publicly available from the Portuguese National Institute of Statistics. The available information is divided into several themes, like population, justice, education, health, or the environment, among several others. The following datasets were specifically used in this case study, in all cases using data for the year of 2011 (i.e., the year of the last national census study):

- Number of female residents, according to the national census in 2011;
- Number of live births in 2011, by place of residence of the mother;
- Number of deaths in 2011, according to the national directorate-general of health;
- Number of foreign residents, according to the national census in 2011;
- Number of buildings, according to the national census in 2011;
- Number of buildings with at least two floors, according to the national census in 2011;
- Resident population employed in the agriculture, animal production, hunting, forest, and fishery sectors, according to the national census in 2011;
- Employed resident population, according to the national census in 2011;
- Number of crimes registered by the police forces in 2011;
- Number of hotel visitors (i.e., number of guests in hotel establishments) in 2011, according to the national tourism authority.

Table 4.1 presents aggregated information for all ten variables listed in the previous enumeration, considering large territorial divisions corresponding to NUTS III regions as the aggregation units. For testing the spatial disaggregation procedure, the 4260 Portuguese civil parishes were first used as the source units, producing raster datasets with a resolution of 30 arc-seconds per cell (i.e., the same resolution used in the GPW dataset that was used as ancillary data). Exceptions to this procedure are the two last variables from the previous enumeration (i.e., the number of crimes and the number of hotel visitors), for which one only had access to data aggregated at the level of municipalities.

In order to geo-reference the statistical information into polygons representing each of the Portuguese administrative subdivisions, a shapefile (i.e., a popular format for geographic information system information) containing the Portuguese official administrative map at the year of 2011 was acquired from the Portuguese national institute responsible for the territorial ordnance survey[1]. To obtain a shapefile with the Portuguese municipality divisions, all the polygons for civil parishes with the same municipality were aggregated.

Figure 4.1 presents a grid with multiple choropleth maps (i.e., five maps per row), illustrating the aggregated information at the level of civil parishes, for the considered socio-economic

[1] `http://www.dgterritorio.pt/cartografia_e_geodesia/cartografia/carta_administrativa_oficial_de_portugal__caop_`

Region	Civil Parishes	Female Residents	Live Births	Deaths	Foreign Residents	Buildings	Tall Buildings	Primary Sector Workers	Employed Population	Crimes	Hotel Visitors
Minho-Lima	290	130,467	1,730	2,834	4,113	120,886	87,603	3,582	91,794	8,655	162,466
Cávado	265	213,346	3,812	2,885	6,932	124,414	91,776	4,263	177,601	13,516	276,779
Ave	243	264,710	4,376	3,793	5,271	157,558	108,765	2,557	217,331	13,844	174,927
Grande Porto	130	676,827	11,798	10,725	23,042	273,491	175,837	6,966	532,190	49,910	1,548,085
Tâmega	321	282,419	4,882	4,063	3,413	197,914	149,181	5,679	219,649	15,455	90,679
Entre Douro e Vouga	80	142,075	2,302	2,163	3,436	89,030	63,903	1,635	119,969	7,909	56,596
Douro	301	107,458	1,429	2,433	2,388	119,398	89,851	10,616	74,908	5,766	142,488
Alto Trás-os-Montes	398	106,120	1,196	2,682	2,882	127,220	96,008	7,725	68,441	7,256	189,957
Baixo Vouga	114	203,744	3,235	3,701	8,872	149,921	82,376	4,398	168,834	14,563	269,109
Baixo Mondego	119	175,723	2,732	3,591	6,547	128,139	73,083	3,601	139,188	11,814	402,626
Pinhal Litoral	66	135,066	2,300	2,345	8,167	109,618	52,372	2,231	113,204	9,004	168,198
Pinhal Interior Norte	115	68,851	861	1,529	3,142	85,699	66,110	1,610	48,737	3,659	38,098
Dão-Lafões	223	145,686	2,153	3,139	4,353	145,974	110,753	5,050	104,755	7,444	236,080
Pinhal Interior Sul	43	21,384	202	693	521	30,618	20,885	820	13,522	850	18,693
Serra da Estrela	67	23,128	256	650	503	28,969	24,687	664	14,867	1,168	32,079
Beira Interior Norte	239	54,859	660	1,529	1,347	74,429	51,421	2,654	37,693	2,681	96,258
Beira Interior Sul	58	39,342	547	1,171	1,320	46,039	29,845	1,226	27,915	2,412	60,909
Cova da Beira	67	45,844	606	1,128	1,134	44,461	34,175	1,272	32,789	2,309	160,474
Oeste	121	187,423	3,188	4,008	14,212	160,794	76,453	9,497	152,172	13,557	317,017
Médio Tejo	103	115,442	1,602	2,613	4,722	107,291	48,688	1,995	86,535	6,497	417,669
Grande Lisboa	153	1,081,345	22,761	18,067	159,517	277,387	196,235	3,992	898,041	98,515	3,683,471
Península de Setúbal	58	405,926	8,366	7,241	44,036	171,570	93,650	4,818	325,235	33,662	341,801
Alentejo Litoral	41	49,301	833	1,299	5,464	53,482	14,824	4,702	40,287	3,064	157,569
Alto Alentejo	86	61,614	881	1,897	2,759	68,275	28,545	3,809	42,554	3,500	132,505
Alentejo Central	91	86,600	1,361	2,016	3,553	80,100	25,528	6,451	67,996	4,406	270,659
Baixo Alentejo	83	64,743	1,016	2,006	3,028	74,901	12,701	5,799	47,217	3,245	96,961
Lezíria do Tejo	91	128,305	2,055	2,889	8,760	107,108	28,695	7,301	100,637	10,193	60,676
Algarve	84	231,075	4,561	4,619	52,046	198,924	96,633	6,142	186,191	25,846	3,008,494
Açores	156	125,238	2,748	2,375	3,356	98,818	56,473	8,636	102,127	10,287	344,595
Madeira	54	141,517	2,407	2,481	5,623	91,961	61,630	3,695	108,808	7,379	1,036,864
Overall	4260	5,515,578	96,993	103,203	394,459	3,544,389	2,148,686	133,386	4,361,187	415,325	13,992,782

Table 4.1: The socio-economic variables considered in the Portuguese case study.

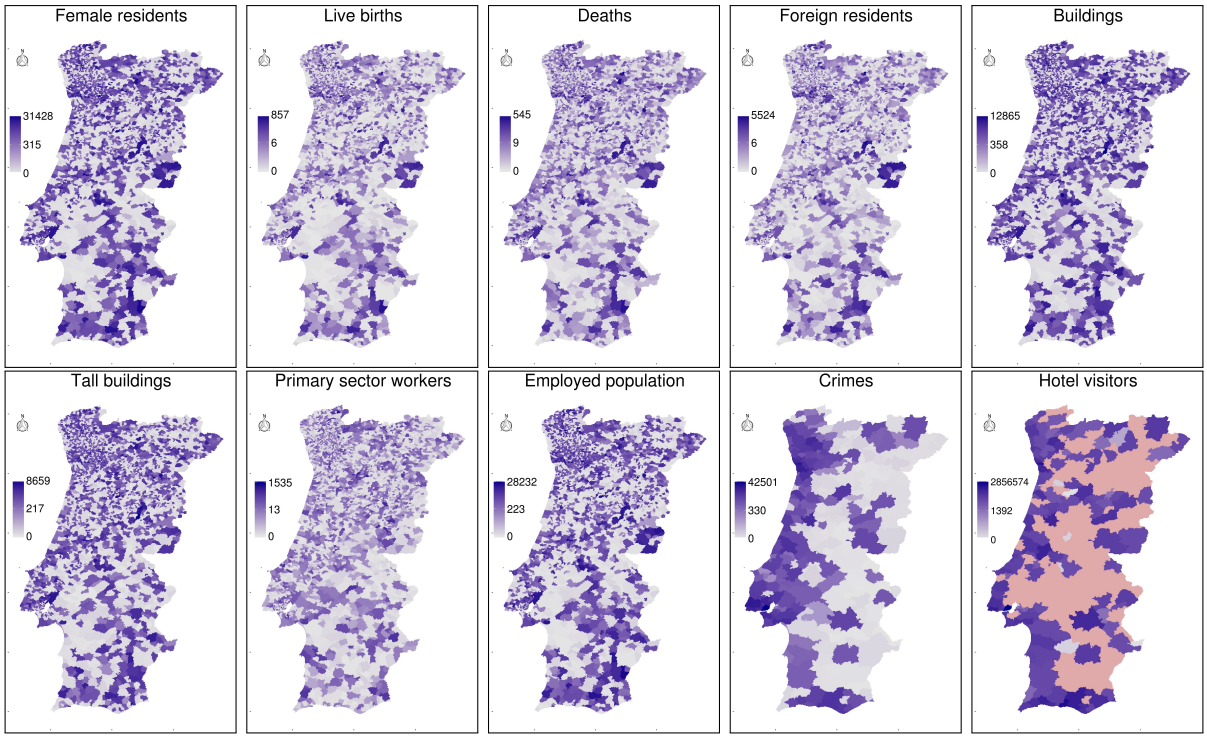

Figure 4.1: Aggregated data for the different socio-economic indicators.

indicators and for the administrative units in Continental Portugal (i.e., ignoring the archipelagos of Azores and Madeira). In the case of the variables corresponding to (i) the number of crimes registered by the police forces, and (ii) the number of hotel visitors, the maps displayed on Figure 4.1 use municipalities as the aggregation level, instead of civil parishes. Information on the number of hotel visitors is also available only for some of the municipalities in the Portuguese territory. Thus, in the corresponding map, the regions shown in red correspond to those municipalities where no information was available. All the maps from Figure 4.1 used a logarithmic transformation to assign data values to particular colors, given that most of the indicators that were considered for disaggregation have a skewed distribution in their values.

Figure 4.2 presents a similar grid to the one that is shown in Figure 4.1, but in this case illustrating the results that were obtained through the complete spatial disaggregation procedure that was proposed in this dissertation, using as source zones the polygons at the highest possible resolution in terms of the original data aggregation (i.e., civil parishes, in all cases except for the indicators corresponding to the number of crimes and the number of hotel visitors). The number of crimes was disaggregated from the level of municipalities, and the number of hotel visitors was disaggregated from a NUTS III level, given that these data were available for the entire NUTS regions, although not for some of the municipalities). All sources of ancillary information, except social media data, were also used in this first test, together with linear regression models within the disseveration-based algorithm. The maps from Figure 4.2 have a resolution of 30

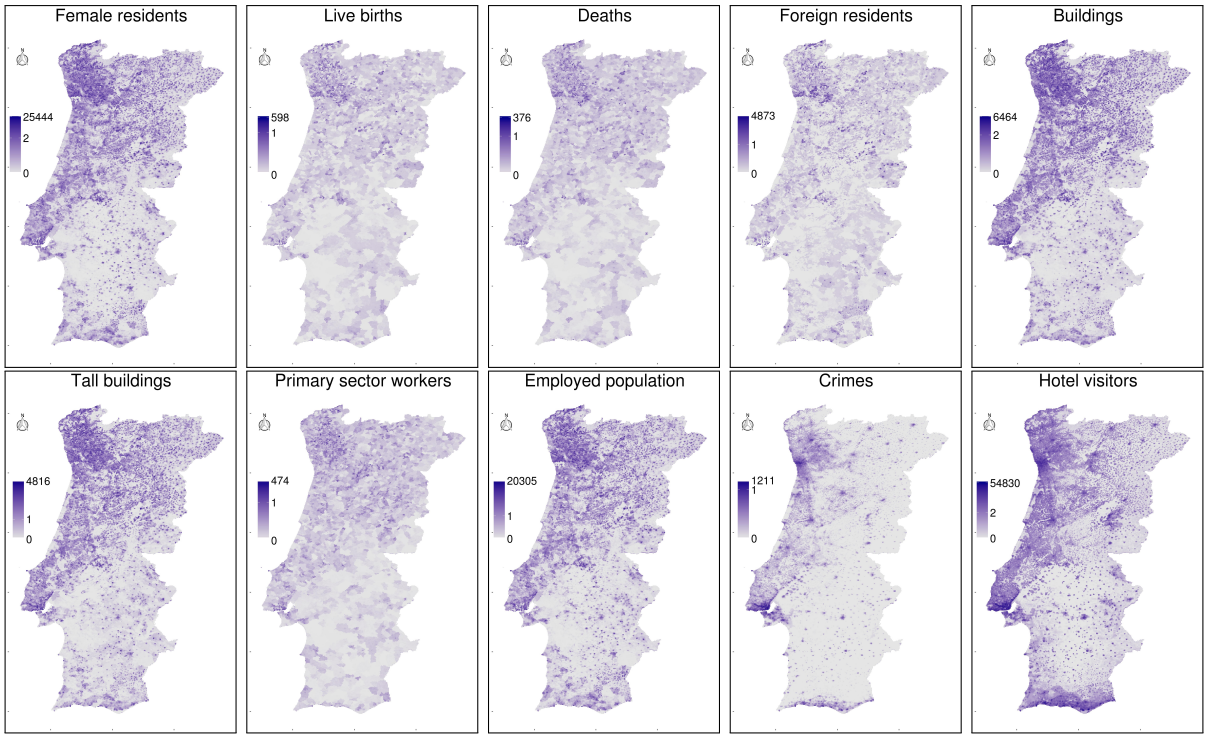

Figure 4.2: Disaggregation results for the different socio-economic indicators.

arc-seconds per cell, and they illustrate general trends in the resulting distribution for the disaggregated values (e.g., higher values are assigned to coastal regions).

Figure 4.3 details one of the variables from Figures 4.1 and 4.2, specifically the one corresponding to the number of crimes. This figure plots, side-by-side, (i) a choropleth map with the number of crimes per municipality, (ii) the ancillary raster with population counts for the Portuguese territory, (iii) a raster showing the disaggregated number of crimes, as obtained with a simpler method corresponding to a proportional and weighted areal interpolation procedure that only used population data (i.e., raster T^d from the enumeration shown in Section 3.1), and (iv) the raster obtained with the proposed hybrid disaggregation method, using linear regression with all sources of ancillary data. From the figure, one can see that indeed the areas with the highest population counts end up receiving a large proportion of the disaggregated counts for the number of crimes, and also that the resulting map is smoother than the one that would be produced by the proportional and weighted areal interpolation procedure.

Figure 4.4 details another of the variables from Figures 4.1 and 4.2, specifically the disaggregated number of foreign residents. This figure plots, side-by-side, (i) a choropleth map with the number of foreign residents per civil parish, (ii) the ancillary raster with nighttime light emissions for the Portuguese territory, (iii) a raster showing the disaggregated number of foreign residents, as obtained with a simpler method that only used population data in the disaggrega-

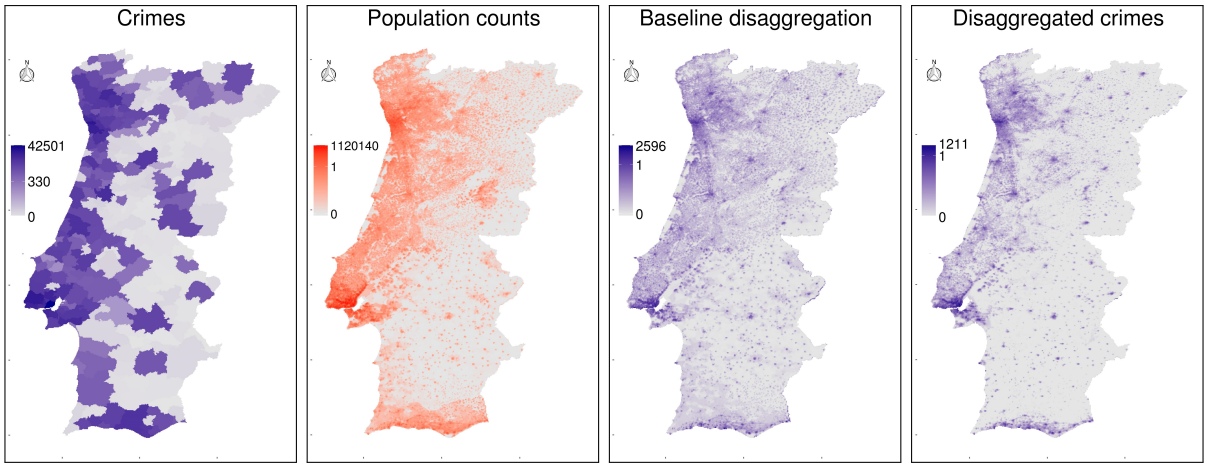

Figure 4.3: Spatially disaggregated results for the variable corresponding to the number of crimes.

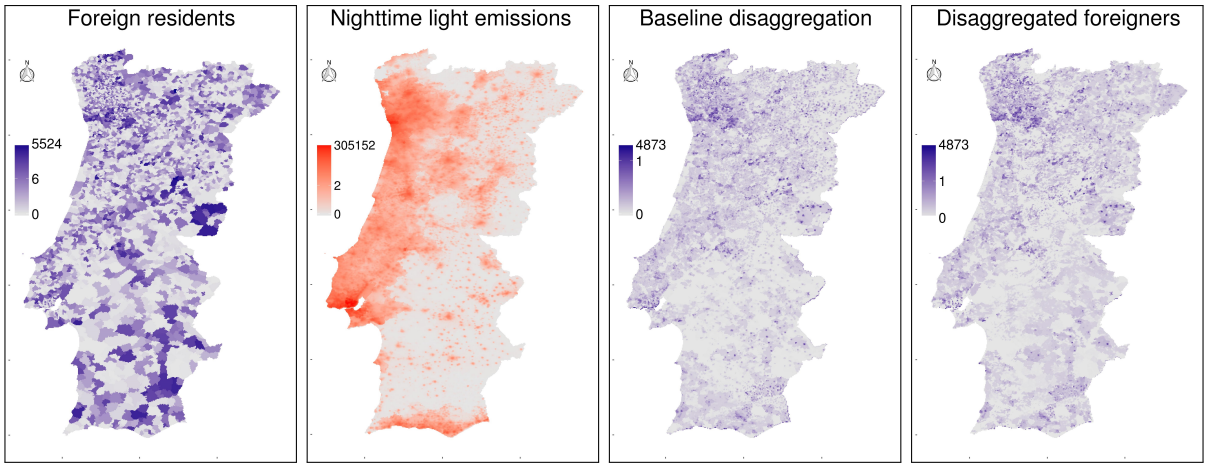

Figure 4.4: Spatially disaggregated results for the variable corresponding to the number of foreign residents.

tion procedure, and (iv) the raster obtained with the proposed hybrid disaggregation method, again using linear regression with all sources of ancillary data, except social media density. From the figure, one can also see that indeed the coastal areas with the highest population counts end up receiving a large proportion of the disaggregated values for the number of foreigners, and also that the nighttime light emission data may aid in the disaggregation procedure, given that light intensity seems to be correlated with the target variable.

In sum, the results from the previous figures suggest that indeed there is a high correlation between variables such as population counts or nighttime light emissions, and the target variables that are to be disaggregated. When investigating these correlations, a very strong linear correlation was found between the population counts for each aggregation area (e.g., for each civil parish) and the aggregated values for the different variables that were considered. Consequently, a very high linear correlation is also found for the disaggregated results produced

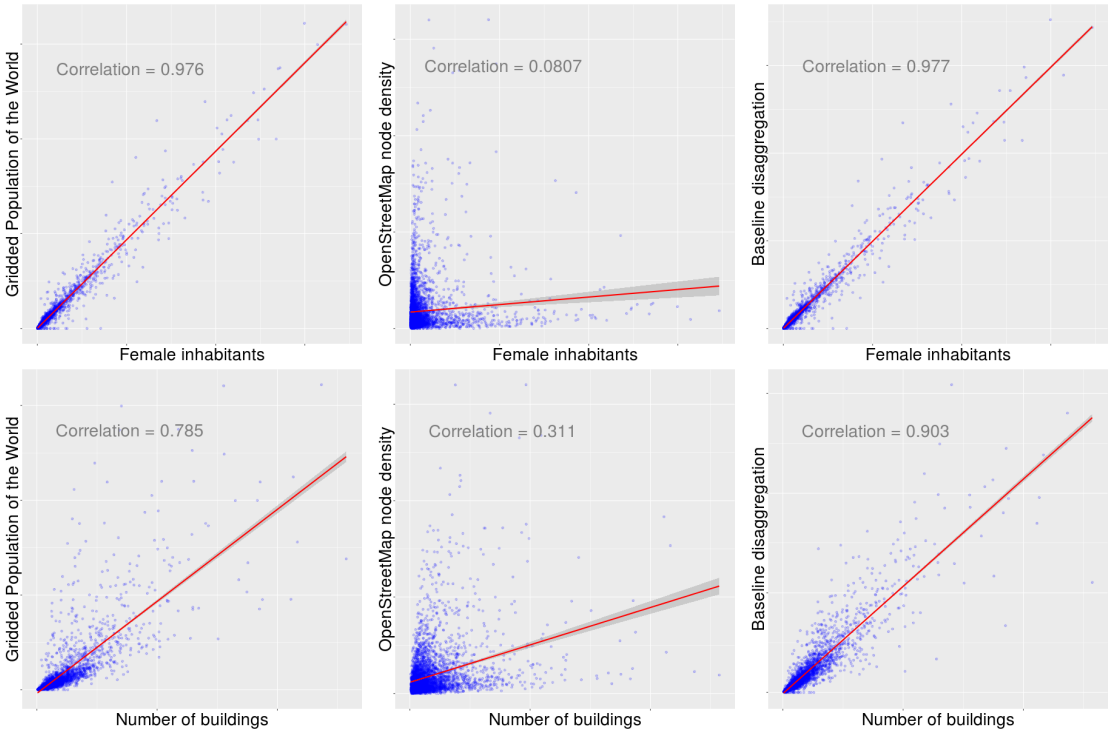

Figure 4.5: Correlations between the variables corresponding to the *number of female inhabitants* (top) and *number of buildings* (bottom), against three of the considered datasets with ancillary information.

through the dasymetric procedure that relied exclusively on population counts, as ancillary information (i.e., results suggest that proportional and weighted areal interpolation, leveraging the population counts, constitutes a very strong baseline).

Figure 4.5 presents scatter-plots illustrating the correlation between two of the target variables that were considered in this case study (i.e., the number of female residents and the number of buildings per civil parish, respectively the variables with the highest/lowest linear correlation towards population counts) with three of the ancillary rasters used in the disaggregation procedure based on disseveration, namely the information regarding the population counts, the OpenStreetMap node count, and the raster obtained with the simple disaggregation method that only used population counts as ancillary information. The three rasters with ancillary information were aggregated to the level of civil parish, in order to compare these values against those from the ground-truth information for the target variables that are to be disaggregated. Each plot presents also the actual value that was obtained for the Pearson correlation coefficient, which in all cases had a p-value below the threshold of 0.001.

To measure the errors that are involved in the proposed spatial disaggregation procedure, concerning the Portuguese case study, the large territorial divisions from which the data were disaggregated were the NUTS III divisions shown in Table 4.1, or the 308 municipalities. The

	Pycnophylactic Interpolation				Weighted Interpolation				Hybrid Method			
	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE
Female residents	3944.16	1379.01	11.387	3.981	**839.76**	**278.52**	**2.425**	**0.804**	840.21	283.43	2.426	0.818
Live births	79.73	26.30	9.303	3.068	108.37	10.72	12.646	1.251	**21.35**	**8.38**	**2.491**	**0.977**
Deaths	68.43	24.16	12.556	4.432	107.48	11.43	19.722	2.098	**19.95**	**9.15**	**3.661**	**1.679**
Foreign residents	517.34	123.31	8.299	1.978	194.51	47.36	3.120	0.760	**164.00**	**44.74**	**2.631**	**0.718**
Buildings	1386.72	631.27	10.806	4.919	786.92	329.22	6.132	2.565	**747.37**	**308.31**	**5.824**	**2.402**
Tall buildings	906.69	427.85	10.471	4.941	497.83	225.28	5.749	2.602	**487.98**	**223.98**	**5.636**	**2.587**
Prim. sect. workers	54.57	24.61	3.555	1.603	120.48	27.28	7.849	1.777	**45.74**	**20.73**	**2.980**	**1.351**
Employed pop.	3215.25	1118.42	10.593	3.685	738.62	**256.98**	2.433	**0.847**	**738.00**	261.35	**2.431**	0.861
Crimes (M)	3295.44	1184.76	7.755	2.788	1298.86	383.84	3.057	0.903	**1248.36**	**346.14**	**2.938**	**0.815**
Hotel visitors (M)	288576.60	98276.07	10.103	3.441	195708.60	**59899.58**	6.852	**2.097**	**195681.50**	60013.58	**6.851**	2.101

Table 4.2: Disaggregation errors measured for the ten different socio-economic variables, with the aggregated data collected originally at a NUTS III level.

aggregated estimates obtained through the disaggregation procedure were then compared against the originally available values for the 4260 civil parishes from 308 municipalities.

Table 4.2 shows the obtained results, in the case of aggregated data collected at the NUTS III level, comparing the usage of the complete hybrid disaggregation method, when leveraging linear regression models, against the results obtained with (i) pycnophylactic interpolation, or with (ii) weighted areal disaggregation leveraging population data for the weights (i.e., raster T^d in the enumeration given in Chapter 3). All the evaluation metrics are computed over the results at the level of civil parishes, except for the last two variables (i.e., the number of crimes and the number of hotel visitors) for which the information at a finer granularity than municipalities was not available. The results for the NRMSE and NMAE metrics are reported with a multiplication factor of 10^{-2}, in order to facilitate the interpretation of quantities associated to small areas. Values in bold correspond to the best results that were achieved for each variable.

Tables 4.3 and 4.4 present similar results, in this case obtained with the disaggregation of data originally reported at the level of municipalities, and measuring results at the level of civil parishes. Table 4.3 presents results for baseline methods corresponding to (i) mass-preserving areal weighting, (ii) pycnophylactic interpolation, and (iii) weighted areal disaggregation leveraging population data for the weights. Table 4.4 instead presents results with the hybrid disaggregation method based on disseveration, using three different regression methods in the dasymetric procedure, namely linear regression models, generalized additive models, or ensembles of trees

	Areal Interpolation				Pycnophylactic Interpolation				Weighted Interpolation			
	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE
Female residents	2220.60	876.94	6.411	2.532	2097.99	845.73	6.057	2.442	**587.44**	**185.57**	**1.696**	**0.536**
Live births	47.21	17.40	5.509	2.031	44.72	16.74	5.218	1.953	15.12	6.07	1.764	0.708
Deaths	35.98	15.57	6.601	2.857	35.61	15.33	6.535	2.813	**15.76**	**6.83**	**2.893**	**1.253**
Foreign residents	322.70	82.41	5.176	1.322	326.06	81.55	5.230	1.308	133.91	35.35	2.148	0.567
Buildings	786.00	412.58	6.125	3.215	793.52	417.58	6.183	3.254	538.06	241.46	4.193	1.882
Tall buildings	548.98	290.48	6.340	3.355	554.03	291.93	6.398	3.371	317.16	155.76	3.663	1.799
Primary sector	41.83	16.97	2.725	1.105	44.72	17.89	2.913	1.165	37.42	16.92	2.438	1.102
Employees	1871.29	724.62	6.165	2.387	1761.49	696.99	5.803	2.296	506.61	172.12	1.669	0.567

Table 4.3: Disaggregation errors measured for different socio-economic variables, using baseline methods and with the aggregated data collected originally at the level of municipalities.

	Linear Models				Generalized Additive Models				Cubist			
	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE
Female residents	589.04	188.44	1.701	0.544	600.76	194.92	1.735	0.563	702.99	237.05	2.030	0.684
Live births	14.99	**5.97**	1.749	**0.696**	**14.91**	6.02	**1.739**	0.702	17.57	6.71	2.050	0.783
Deaths	16.55	7.40	3.036	1.358	17.18	7.40	3.152	1.357	17.72	7.38	3.251	1.355
Foreign residents	133.31	**34.78**	2.138	**0.558**	**133.21**	36.66	**2.137**	0.588	180.71	45.36	2.899	0.728
Buildings	511.02	223.21	3.982	1.739	497.11	219.90	3.874	1.714	**329.51**	**170.15**	**2.568**	**1.326**
Tall buildings	311.83	154.74	3.601	1.787	304.25	150.93	3.514	1.743	**263.99**	**138.36**	**3.049**	**1.598**
Prim. sect. workers	37.37	**15.33**	2.435	**0.998**	37.98	16.48	2.474	1.074	**35.33**	15.99	**2.302**	1.042
Employed pop.	**503.83**	**169.80**	**1.660**	**0.559**	546.68	201.98	1.801	0.665	622.32	230.87	2.050	0.761

Table 4.4: Disaggregation errors measured for different socio-economic variables, using different types of regression models and with the aggregated data collected at the level of municipalities.

based on the cubist method. The results for the NRMSE and NMAE metrics are again reported with a multiplication factor of 10^{-2}. The values shown in bold, in both Tables 4.3 and 4.4, correspond to the best results that were achieved for each variable (i.e., values in bold shown in Table 4.3 correspond to cases in which the hybrid disaggregation method, based on dissever-ation, could not outperform one of the baselines, namely the one based on weighted areal disaggregation leveraging population).

The results from Table 4.2 show that the proposed hybrid method indeed outperforms the baselines corresponding to pycnophylactic interpolation or weighted areal interpolation, at a NUTS III level. However, in some error metrics and particularly for indicators that have a strong linear correlation towards population counts (e.g., the indicator corresponding to the number of female residents), the simpler dasymetric procedure that only takes into account the population as ancillary data produces slightly better results.

From Tables 4.3 and 4.4 one can also see that, at a municipality level, the hybrid method continues to outperform the baseline disaggregation methods in almost all indicators. When the indicator to disaggregate is strongly correlated with population counts (e.g., for variables such as female residents, live births, or employed population), the methods that produced lower disaggregation errors used regression analysis based on standard linear regression or generalized additive models. The strong linear dependence between the indicators that are to be disaggregated and some of the ancillary variables can explain why a simple linear regression can model the dependence better than more sophisticated methods. On the other hand, for the case of indicators depending less on population (e.g., number of buildings, or number of buildings with more than a single floor), the regression model based on ensembles of trees obtained slightly better results. In all cases, the simple method based on weighted areal disaggregation, leveraging population data for the weights, indeed corresponded to a very strong baseline.

The strong correlations between the considered indicators and the auxiliary variable corresponding to population counts explain why the simpler baseline achieves almost the same results

as the more sophisticated hybrid method. Nonetheless, for almost all indicators, the usage of additional ancillary information can indeed lead to improvements, sometimes considerable ones. It should also be noted that the errors that were reported correspond to an upper bound on the actual errors produced from the disaggregation of data reported at the level of civil parishes (i.e., these tests have only measured the errors in the disaggregation of data originally at the level of NUTS III regions or municipalities), given that the higher the differences between the source and the target areas, the higher the errors introduced by a disaggregation procedure.

It is also interesting to notice that the errors that were measured for the different spatial disaggregation procedures were also evenly distributed over the considered geographic territory. In Figure 4.6, the errors that were measured individually for each civil parish (i.e., the difference between the estimated value and the real value, divided by the real value so as to obtain normalized scores) are plotted, in the case of the indicators that generally had the lowest and highest errors (i.e., the number of female residents and the number of buildings, respectively) over the continental Portuguese territory, and for the archipelagos of Azores and Madeira. These errors are shown for the case of the disaggregation procedures corresponding to (i) the disseveration algorithm based on a linear model, and (ii) the weighted areal interpolation procedure that leveraged population data. From the figures, one can see that the largest errors in both disaggregation procedures correspond to civil parishes for which the true values were over-estimated. The regions with darker colours and higher values generally correspond to civil parishes where the indicator being disaggregated had very small values (i.e., a number of 268 female residents in the civil parish named *São Nicolau*, in the municipality of *Mesão Frio* and district of *Vila Real*, over continental Portugal), and where therefore a small deviation in the estimated results produced high values for the normalized error. Still, the disaggregation errors are very small in most of the territory, and also evenly distributed.

4.3 Experiments with Tourism Indicators

Using the data that were also considered for the experiments reported on the previous section, I first attempted to validate the hypothesis that ancillary data extracted from popular location-based services, like Flickr, highly correlates with the distribution of variables related to tourism. The Flickr dataset reported in Section 3.3 was thus also leveraged to disaggregate the number of hotel visitors in the Portuguese territory, but this variable had almost no impact on the results, given that a large portion of the Portuguese territory had no Flickr photos. As the number of Flickr photos has a higher density in big cities, such as Lisbon or Porto, the study

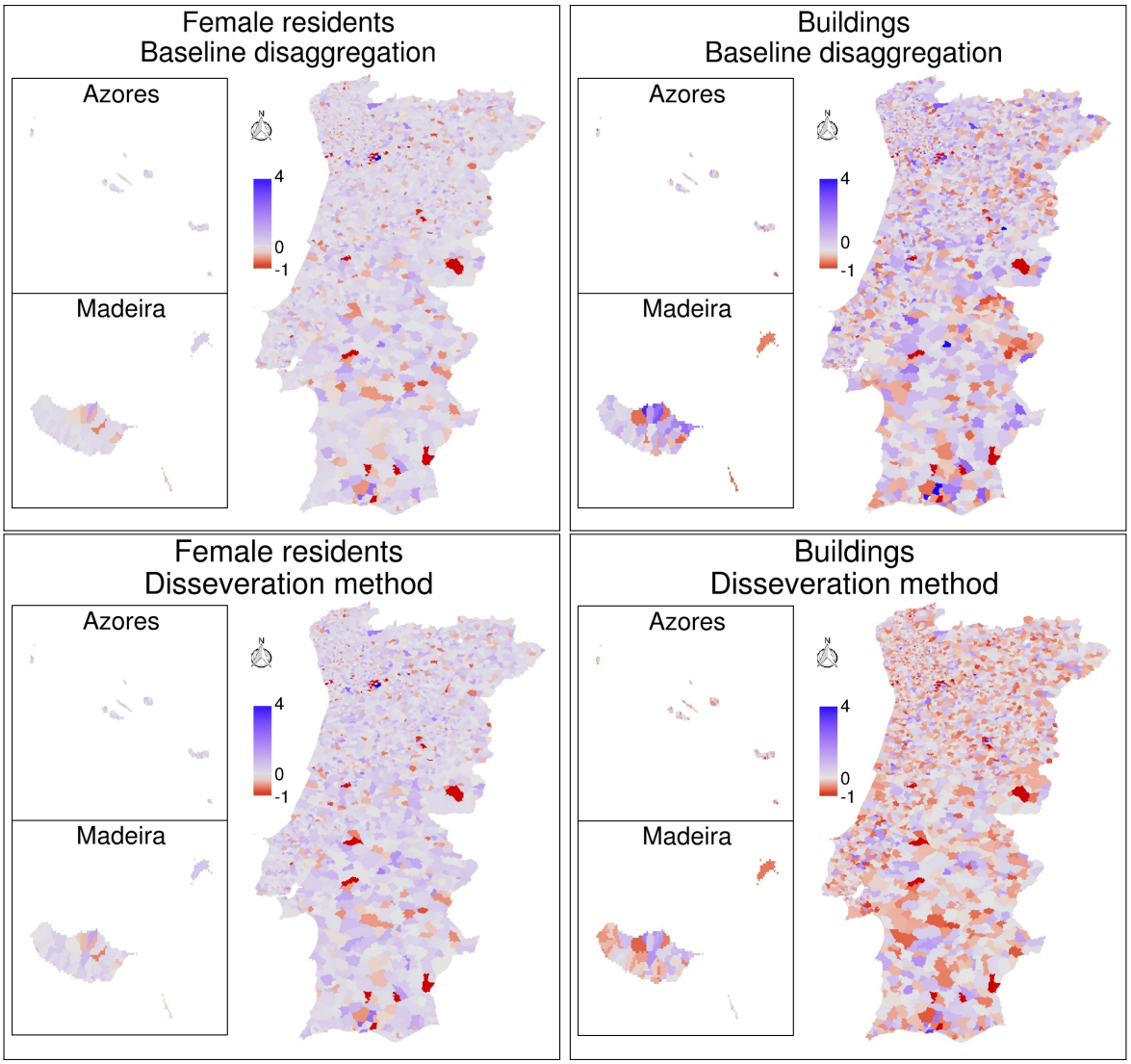

Figure 4.6: Normalized errors measured for the different civil parishes.

region was redefined to the area surrounding the Portuguese capital, i.e., the NUTS III region of *Grande Lisboa*. Two experiments were then carried out, one leveraging only the baseline ancillary information, i.e., the datasets used in the Portuguese study case, and another one also leveraging the Flickr dataset. The resolution of the ancillary information was also redefined to 15 arc-sec, since the region size was much smaller than in earlier tests. Several experiments were performed, using different values for the *bandwidth* parameter of the kernel density estimation method, used to compute the density of the Flickr photos. Table 4.5 reports the measured errors for the disaggregation of the number of hotel visitors, when setting the *bandwidth* to 0.02, since this value has produced the best results in these tests.

A closer look at the data presented in Table 4.5 indicates that the best results are achieved with a linear regression model or geographically weighted regression (i.e., depending on the error metric), when leveraging only the baseline ancillary information, and with the geographically

	Baseline Datasets				Baseline Datasets + Flickr			
	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE
Linear Models	**1155547.0**	718729.1	**41.006**	25.505	1156956.0	719362.8	41.056	25.527
Robust Linear Model	1157762.0	718744.4	41.084	25.506	1157763.0	**718742.3**	41.085	**25.505**
Generalized Additive Models	1155744.0	718816.9	41.013	25.508	1156909.0	719251.0	41.054	25.523
Cubist	1156561.0	719025.5	41.042	25.515	1156847.0	719219.4	41.052	25.522
Geographically Weighted Regression	1155636.0	**718377.2**	41.009	**25.492**	**1155328.0**	718866.5	**40.998**	25.510

Table 4.5: Disaggregation errors measured for the number of hotel visitors in Lisbon, using different types of regression models and with the aggregated data collected at a NUTS III level.

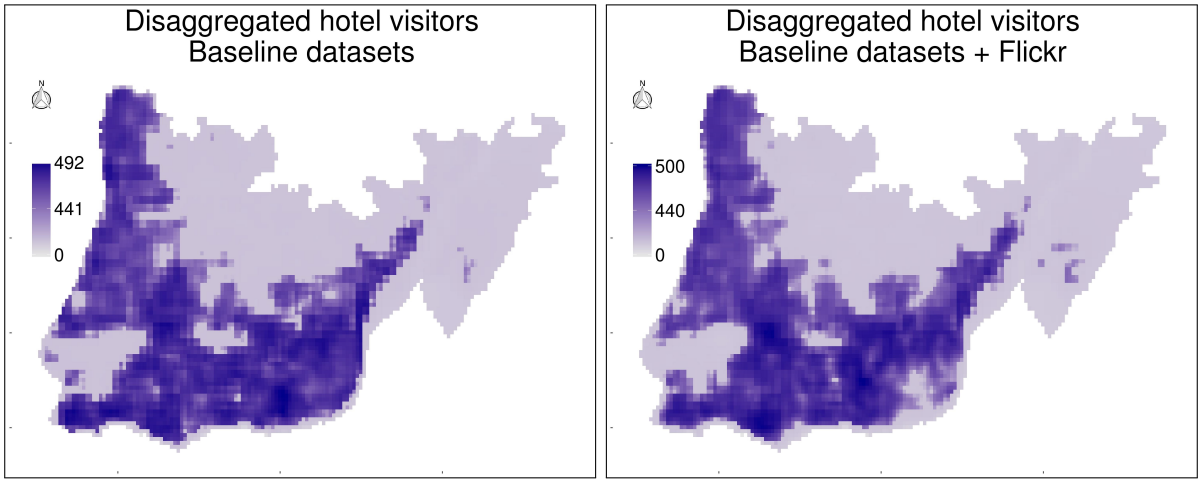

Figure 4.7: Spatially disaggregated results for the number of hotel visitors in *Grande Lisboa*.

weighted regression or a robust linear model, when leveraging also the Flickr dataset. However, the measured errors are quite high, especially when compared with previous experiments on the entire Portuguese territory. The main reason for the high error values is the reduced number of municipalities with available data for the number of hotel visitors in *Grande Lisboa*, i.e., only six different municipalities. With such a small number of divisions for which the aggregated data were going to be compared against the original values, it is almost impossible to get a sense of how the introduction of extra ancillary information may cause benefits in the disaggregation procedure. For illustration purposes, Figure 4.7 presents two rasters with the results of disaggregating the number of hotel visitors in *Grande Lisboa*, when using standard linear regression to leverage only the baseline datasets, or when leveraging the baseline datasets together with the Flickr photo density.

To overcome the aforementioned limitation and build a scenario that allowed me to properly validate the introduction of social media usage as valuable ancillary information, two other case studies concerning the territories of Belgium and France were considered. Two datasets were collected from tourism statistics portals, namely one with data pertaining to three provinces in Belgium, and another with data for the French territory. Taking into account that the available information was dispersed through several data sources (i.e., in different PDF and

Province	Municipality	Overnights	Total
Brussels	Anderlecht	183,987	5,907,382
	Auderghem - Oudergem	21,195	
	Ville de Bruxelles	3,231,420	
	Etterbeek	130,768	
	Evere	132,139	
	Forest	12,912	
	Ganshoren	11,087	
	Ixelles	463,490	
	Jette	12,681	
	Saint-Gilles	629,360	
	Saint-Josse-ten-Noode	726,135	
	Schaerbeek	179,586	
	Uccle	41,258	
	Watermael-Boitsfort	6,214	
	Woluwe-Saint-Lambert	102,342	
	Woluwe-Saint-Pierre	22,808	
Flemish Brabant	...	...	1,692,894
Waloon Brabant	Grez-Doiceau	313	323,876
	Ittre	10,417	
	Lasne	7,742	
	Nivelles	35,224	
	Ottignies-Louvain-la-Neuve	46,727	
	Rixensart	46,180	
	Villers-la-Ville	541	
	Waterloo	85,352	
	Wavre	91,380	

Table 4.6: The number of tourist overnights for the administrative divisions of Belgium.

Excel documents), the information was first represented in a format that was easy to use, namely ESRI Shapefiles, so that the data could be read and processed adequately. The following data, referring to the year of 2012, were used in these case studies:

- Number of overnights in 2012, according to the annual report of the observatory for tourism in Brussels[2];
- Number of nights spent at tourist accommodation establishments in 2012 for the territory of France, according to the Eurostat regional tourism statistics[3].

Table 4.6 presents the collected aggregated information for the first indicator. In this case, the divisions considered in the disaggregation procedure were three provinces, concerning the

[2] http://visit.brussels/site/binaries/content/assets/pdf/report_2012.pdf
[3] http://ec.europa.eu/eurostat/web/regions/data/main-tables

NUT I	NUT II	Nights spent	Total
Ile de France	Ile de France	36,215,571	36,215,571
Bassin Parisien	Champagne-Ardenne	2,520,825	27,359,053
	Picardie	3,748,863	
	Haute-Normandie	3,806,268	
	Centre	6,808,277	
	Basse-Normandie	6,262,583	
	Bourgogne	4,212,237	
Nord - Pas-de-Calais	Nord - Pas-de-Calais	5,376,125	5,376,125
Est	Lorraine	4,578,780	12,532,971
	Alsace	5,144,749	
	Franche-Comte	2,809,442	
Ouest	Pays de la Loire	16,455,836	43,419,111
	Bretagne	15,166,836	
	Poitou-Charentes	11,796,439	
Sud-Ouest	Aquitaine	23,241,615	38,457,925
	Midi-Pyrenees	13,033,780	
	Limousin	2,182,530	
Centre-Est	Rhone-Alpes	34,554,724	40,249,329
	Auvergne	5,694,605	
Mediterranee	Languedoc-Roussillon	25,345,917	68,594,425
	Provence-Alpes-Cote d'Azur	37,030,502	
	Corse	6,218,006	

Table 4.7: The number of visitors for the administrative divisions of France.

region of Brussels and its outskirts, that are sub-divided into 59 municipalities. On the other hand, Table 4.7 introduces the collected aggregated information for the second indicator. The divisions considered were, in this case, the 8 French NUTS I divisions (i.e., all the NUTS I regions except *Départements d'Outre Mer*), that are sub-divided into 22 NUTS II divisions. The shapefiles used to geo-reference the collected information into polygons representing each of the administrative subdivisions were obtained from the OSM Boundaries Map project[4] (concerning the Belgium provinces) and from the Eurostat portal[5] (concerning the French territory).

Figure 4.9 plots, side-by-side, (i) the ancillary raster with nighttime light emissions for the three provinces in Belgium, (ii) the raster containing the density of Flickr photos for the same region, obtained with a kernel density estimation method, (iii) a choroplet map with the number of overnights per municipality, and (iv) a raster showing the disaggregated number of overnights, using a linear regression model with all sources of ancillary data. The regions shown in light brown correspond to municipalities where no information was available. From the figure, one can see that most of the areas that have higher values within the nighttime light emissions and the Flickr density rasters receive a large proportion of the disaggregated counts for the number

[4] `http://osm.wno-edv-service.de/boundaries`
[5] `http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data`

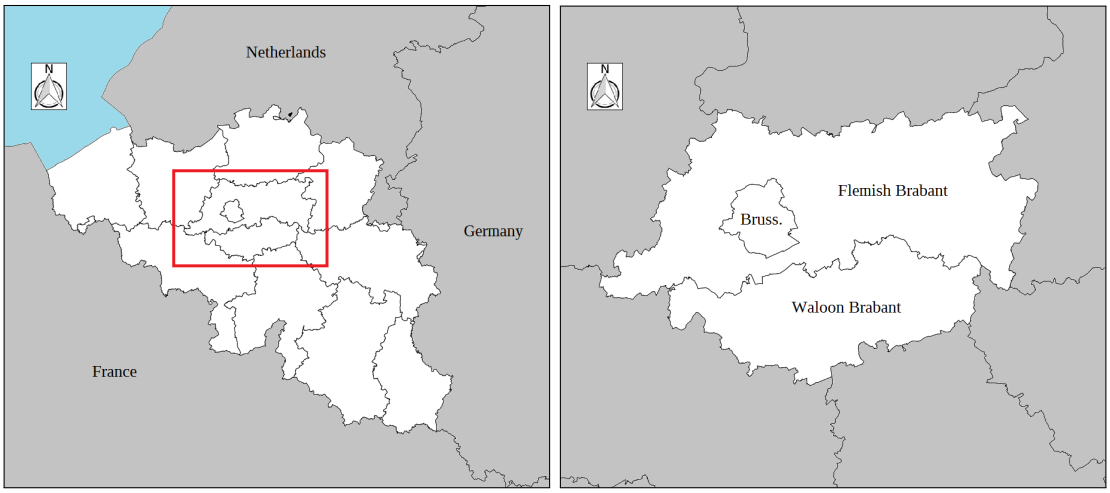

Figure 4.8: Map of the three provinces in Belgium.

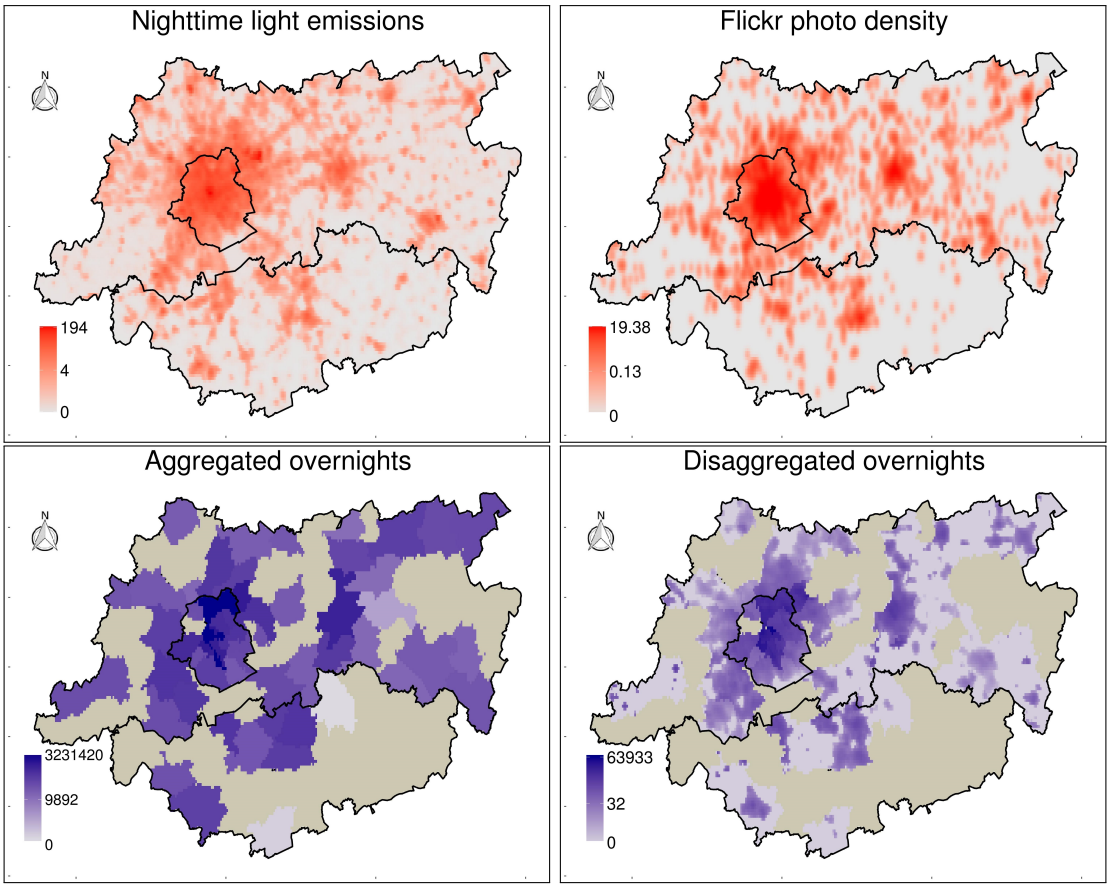

Figure 4.9: Spatially disaggregated results for the number of tourist overnights over the territory of Belgium, together with the source data and the ancillary variables.

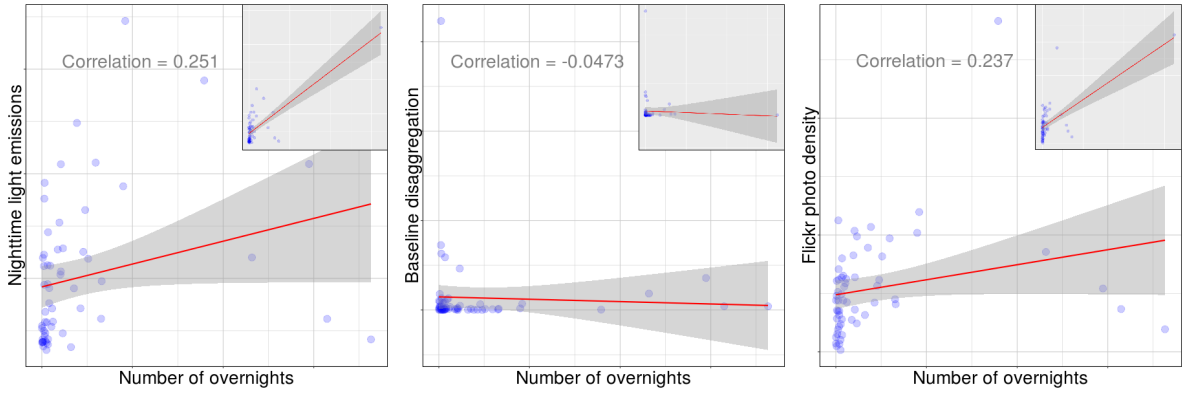

Figure 4.10: Correlation between the variable *number of overnights* and the considered datasets with ancillary information.

of overnights. However, one can also see that the correlation between the indicator involved in the disaggregation and these ancillary variables is not as high as in previous experiments (e.g., in the Portuguese case study, in which the considered variables were much more correlated with the ancillary information). In any case, the map with the disaggregated overnights shows a clear distinction between high and low values inside each of the municipalities under study.

The correlation between the aggregated values for the number of overnights against the aggregated values for each of the ancillary datasets, at a municipality level, is shown through scatter-plots in Figure 4.10. Each scatter-plot was computed after removing the values relative to the region of *Ville de Bruxelles*, which constitutes an outlier due to the fact that it has a very disproportional value for the indicator under study. However, an inset with the same scatter-plot, using all the data values (i.e., including the outlier), is also presented in each plot, for illustration purposes.

From the figure, one can see that the two ancillary datasets that have higher correlations with the indicator under study are the ones regarding the nighttime lights and the Flickr photo densities. It is thus expected that the disaggregation procedure achieves better results when leveraging these two datasets. Additionally, it is also expected that a more sophisticated regression model, that seeks for non-linear correlations between the indicator and the ancillary information, will tend to produce lower error values, since there are no visible strong linear correlations illustrated in the scatter-plots from the figure.

Table 4.8 shows the results for the measured errors in the case of the disaggregation of the number of overnights in Belgium, that were obtained by comparing the aggregated estimates for the 59 municipalities that overlapped with the referenced provinces, against the original values. Two experiments are reported, namely one leveraging nighttime lights and the computed raster with the baseline disaggregation as ancillary information (since these rasters are the only ones

	Nighttime Lights + Baseline Disaggregation				Nighttime Lights + Baseline Disaggregation + Flickr			
	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE
Linear Models	241229.8	114410.8	7.466	3.541	251930.5	115966.4	7.797	3.589
Robust Linear Model	243911.5	115124.9	7.549	3.563	256355.6	119296.5	7.934	3.692
Generalized Additive Models	**215938.8**	**106209.3**	**6.683**	**3.287**	**208940.6**	**25072.2**	**6.467**	**2.942**
Cubist	309153.3	147551.6	9.568	4.567	317815.0	148054.7	9.836	4.582
Geographically Weighted Regression	327534.8	142686.8	10.137	4.416	320108.1	142837.4	9.907	4.420

Table 4.8: Disaggregation errors measured for the number of overnights in the 59 Belgian municipalities, using different types of regression models and with the aggregated data collected at the level of provinces.

that are available at the higher resolution without needing any type of interpolation), and another one leveraging also the Flickr dataset, as explained in Section 3.3. The error results were compared when using the following regression models in the disaggregation method: (i) a simple linear model, (ii) a robust linear model, (iii) a generalized additive model, (iv) the cubist approach, and (v) geographically weighted regression. The values in bold correspond to the best results achieved for each experiment, and the results for the NRMSE and NMAE metrics are again reported with a multiplication factor of 10^{-2}, to make them more easy to read.

From Table 4.8, one can see that the generalized additive model outperforms the other methods in all of the reported metrics, followed by the linear model, when leveraging only the nighttime lights dataset together with the raster containing the baseline disaggregation, and when leveraging also the Flickr dataset. Since generalized linear models constitute more sophisticated regression approaches, and bearing in mind that there is no apparent linear correlation between the indicator and the ancillary datasets, the better results produced by the generalized additive model were, to a certain degree, expected.

When comparing the results obtained with and without the Flickr dataset, using a generalized additive model as the regression approach, one can also see improvements. This model could effectively take advantage of the extra information, producing a much lower error in comparison to the remaining models. On the other hand, the simple linear model, the robust linear model, the cubist approach, and the geographically weighted regression model seem to originate higher errors when leveraging the Flickr dataset.

Figure 4.11 shows the distribution of the normalized error values over the geographic territory, for the disaggregation of the number of overnights in the three Belgian provinces, based on generalized additive models and using all types of ancillary information. Although there are municipalities where the disaggregated indicator had relatively higher values, the errors were evenly distributed over the considered geographic territory.

Focusing on the particular case of the results obtained with geographically weighted regres-

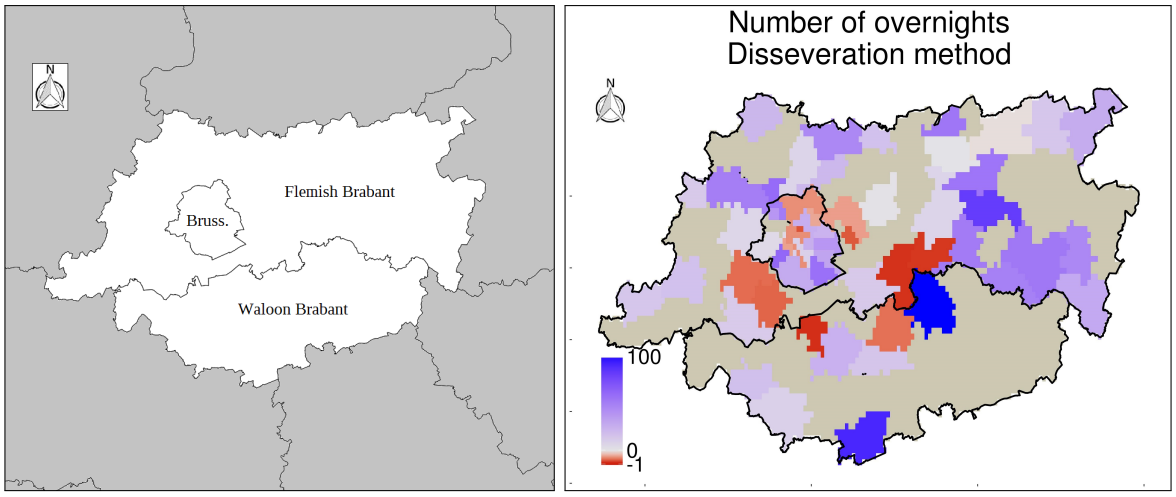

Figure 4.11: Normalized errors measured for the different municipalities.

sion, it is possible to speculate on the possible reasons that make this method inadequate to the previous test case, producing poor results in comparison to the generalized additive model. The geographic area used in this particular case study is perhaps too small to see advantages in exploring multiple regression models, with distinct regional characteristics. Given that the considered divisions are geographically near each other, the distribution of the variable under study is similarly correlated to the ancillary information in all of the regions.

In addition to the Belgian case study, two more experiments were carried out regarding the disaggregation of the number of nights spent by tourists in the French territory, leveraging all types of ancillary information. In particular, the divisions considered as the source zones in the disaggregation algorithm were the 8 NUTS I divisions indicated in Table 4.7 (i.e., excluding the NUTS I region corresponding to *Départements d'Outre Mer*). The obtained aggregated estimates were then compared against the available values for the 23 NUTS II regions that overlap with the referenced NUTS I source zones. The same regression models that were tested in the previous case study were used in this case, but now the resolution of the ancillary rasters was set back to 30 arc-seconds, due to the large dimensions of the territory that was involved.

Figure 4.12 plots, side-by-side, (i) the ancillary raster with the nighttime light density for the territory of France, (ii) the raster containing the density of the Flickr photos, obtained with a kernel density estimation method, (iii) a choroplet map with the number of nights spent by tourists, per NUTS II region, and (iv) a raster showing the disaggregated number of nights spent by tourists, using a linear regression model with all sources of ancillary data. As in previous cases, the two ancillary datasets that are shown in Figure 4.12 seem to have a high relevance in the disaggregation procedure. One can also see that coastal regions of the territory end up having higher estimated values, as well as some urban centers that have more emissions of

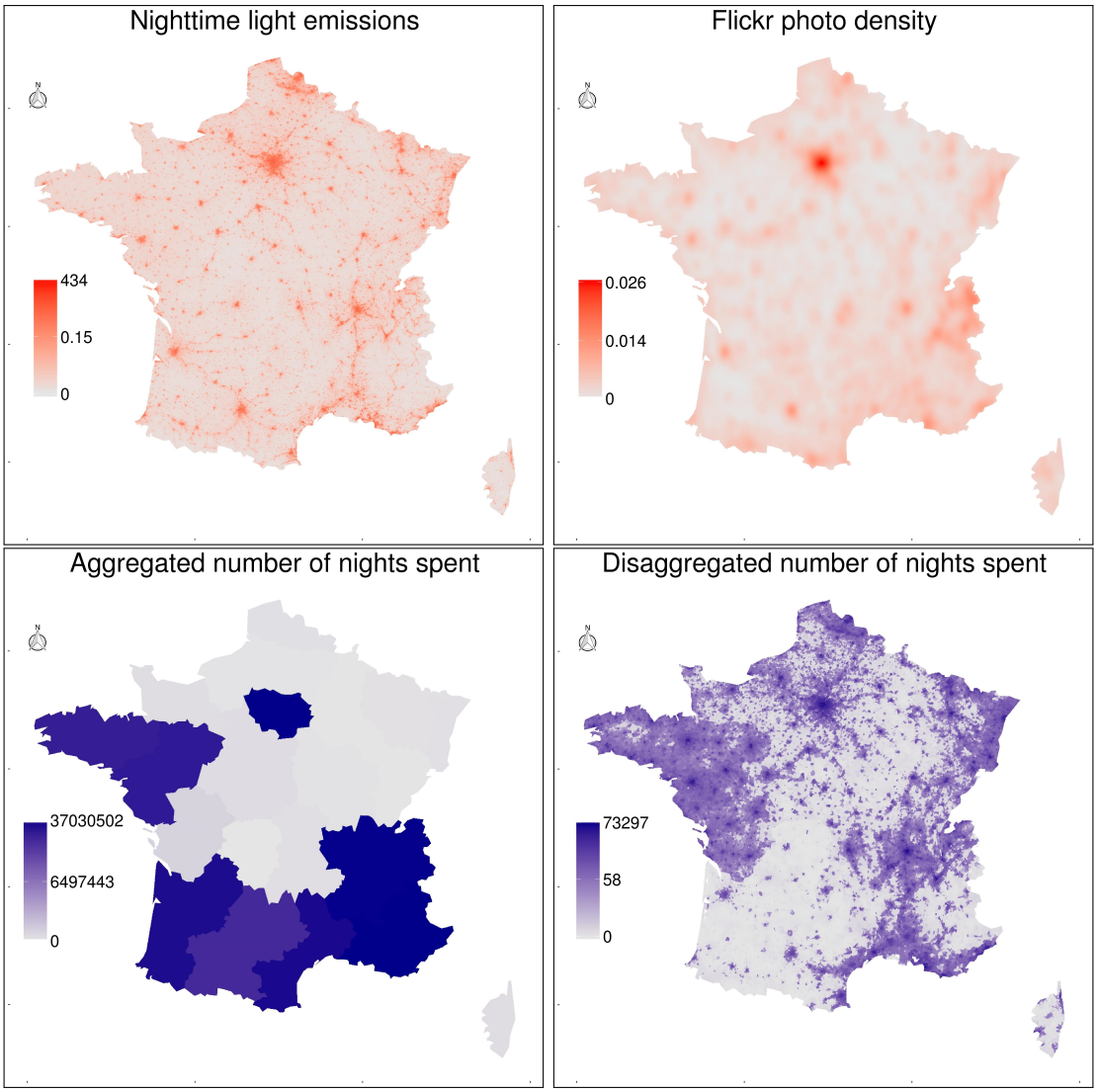

Figure 4.12: Spatially disaggregated results for the number of nights spent over the territory of France, together with the source data and ancillary information.

nighttime lights and a higher density of Flickr photos.

The scatter-plots in Figure 4.13 show the correlation between the aggregated values for the number of nights spent by tourists against each of the ancillary datasets, at a NUTS II level. This figure clearly indicates that the three ancillary datasets have high linear correlation with the indicator under study. Particularly, the datasets concerning with nighttime light emissions and Flickr photo density have the higher values, and so the experiments that leverage these two datasets tend to obtain lower error values.

The obtained disaggregation results, using the aforementioned evaluation methodology, are reported in Table 4.9. Due to the computational effort involved when using geographically weighted regression over the large territory, the number of samples used to calibrate this model was reduced to 0.25% of all the cells in the target raster. Once again, the values in bold

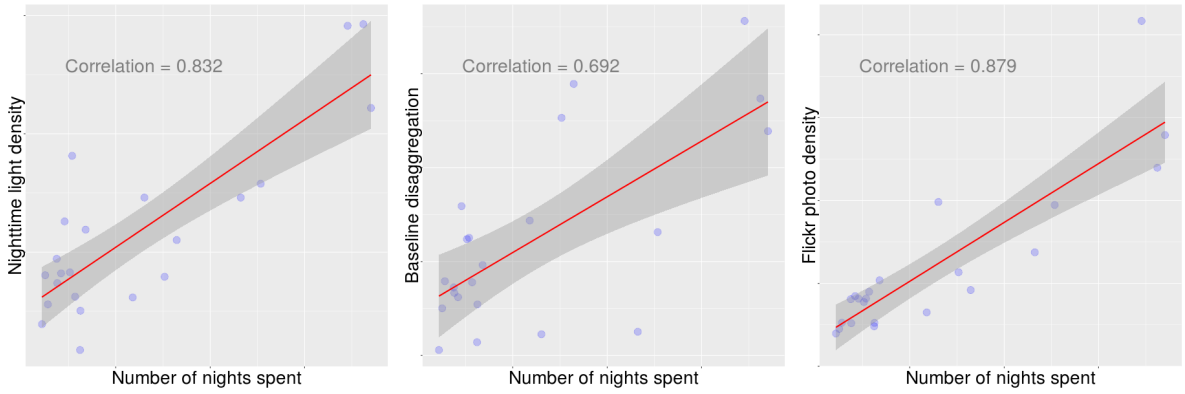

Figure 4.13: Correlations between the variable *number of nights spent by tourists* against the considered datasets with ancillary information.

correspond to the best results achieved for each experiment, and the results for the NRMSE and NMAE metrics are reported with a multiplication factor of 10^{-2}. The reported data suggest that the best results are obtained using the Robust Linear Model.

The distribution of the normalized error values over the French geographic territory, using a robust regression model with all types of ancillary data, is presented in Figure 4.14. From the figure, one can see that the errors are relatively low and evenly distributed over the territory, i.e., there are no neighbouring NUTS II zones whose estimated values were simultaneously too overestimated/underestimated.

4.4 Overview

This chapter presented the experimental evaluation of the novel spatial disaggregation procedure introduced in Chapter 3, describing its application in the disaggregation of different types of indicators. The chapter described the methodology used to perform the different experiments, concerning the distinct geographic territories and the different types of socio-economic indicators. The general procedure that was used to measure the spatial disaggregation errors and to illustrate the application of the technique involved (i) the presentation of the aggregated data for the different geographic regions, (ii) the creation of different disaggregated maps for each of

	Baseline Datasets				Baseline Datasets + Flickr			
	RMSE	MAE	NRMSE	NMAE	RMSE	MAE	NRMSE	NMAE
Linear Models	8346435.0	6612178.0	23.951	18.974	8351701.0	6612227.0	23.966	18.975
Robust Linear Model	**8320851.0**	**6574034.0**	**23.878**	**18.865**	**8346585.0**	**6608128.0**	**23.951**	**18.963**
Generalized Additive Models	8347007.0	6613080.0	23.953	18.977	8354246.0	6611501.0	23.973	18.972
Cubist	8410010.0	6655980.0	24.133	19.100	8362313.0	6670172.0	23.997	19.141
Geographically Weighted Regression	8798823.0	6788310.0	25.249	19.480	8811927.0	6788763.0	25.287	19.481

Table 4.9: Disaggregation errors measured for visitors in France, using different types of regression models and with the aggregated data collected at a NUTS I level.

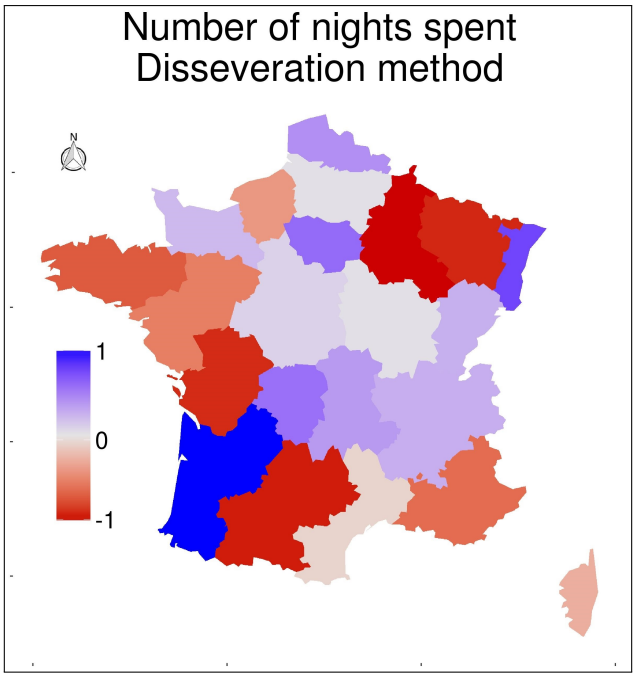

Figure 4.14: Normalized errors measured for the different NUTS II regions.

the indicators, using the disaggregation method, (iii) the analysis of the correlation between the indicators and each of the ancillary variables, (iv) the measurement of the disaggregation errors, and (v) the analysis of the distribution of the errors over the geographic space.

In sum, the disaggregation results that were achieved when applying the novel method outperformed seminal baseline algorithms that were previously used in the literature, such as weighted areal interpolation (Goodchild and Lam, 1980) or pycnophylactic interpolation (Tobler, 1979), in almost all indicators from one of the case studies. The best regression model to combine the multiple sources of ancillary knowledge ranged from the standard linear regression, when disaggregating socio-economic indicators that were strongly correlated with a particular covariate, to more sophisticated methods, like robust regression when leveraging social media usage data in the disaggregation of tourism indicators. Still, the differences between the different types of regression algorithms were always very small, and one can therefore argue that the extra computational effort associated to techniques such as tree ensembles or geographically weighted regression does not compensate in terms of gains in accuracy. Despite the fact that it is not possible to compare the errors measured in the evaluation of the proposed technique with other research work in the area, due to the lack of similar evaluations concerning the same indicators and the same regions, it is my belief that the experiments reported here further validate the benefits of using ancillary data to support disaggregation procedures.

5 Conclusions and Future Work

Spatial analysis in the fields of urban and regional planning, transport planning, or environment impact analysis, often requires high resolution socio-economic data. Such studies typically work with raster data to calculate different types of indicators, like exposure to air pollutants or to noise. However, the available socio-economic data often do not have the necessary spatial resolution. In fact, data on socio-economic variables (e.g., indicators relative to population, employment, public health, or housing) are often available only for large areas, like provinces, districts, municipalities, or other statistical entities. Those units are clearly too coarse to be used in particular types of spatial analysis.

Spatial disaggregation techniques can be used to transform data from a set of source zones into a set of target zones, with different geometry and with a higher spatial resolution. In this dissertation, I reported on experiments with an hybrid spatial disaggregation technique that combines the ideas of dasymetric mapping and pycnophylactic interpolation, using population density, nighttime light emissions, land coverage information, road density data from OpenSteetMap, and social media data extracted from Flickr, as ancillary data to disaggregate different types of socio-economic indicators to a raster-grid level. This combination was carried out using different types of regression models. In fact, the procedure that was implemented can easily use different types of regression approaches, from simple linear regression to state-of-the-art methods like robust or geographically weighted regression. The proposed disaggregation technique was applied in different case studies relative to the Portuguese, Belgian, and French territories, resulting in the production of fine-resolution gridded rasters. In this dissertation, I discussed the spatial disaggregation methodology, as well as the quality of the obtained results. The experiments confirmed that the proposed procedure outperforms seminal baseline algorithms previously used in the literature, such as weighted areal interpolation (Goodchild and Lam, 1980), or pycnophylactic interpolation (Tobler, 1979).

5.1 Overview on the Contributions

The most important contributions of my M.Sc. thesis are as follows:

- **A novel intelligent disaggregation method**: Starting with a downscaling procedure originally outlined by Malone et al. (2012) that uses regression analysis to combine different ancillary variables, this dissertation describes the adaptation/extension of the original procedure for spatial disaggregation, by computing the initial estimates of the dissemination procedure through a mass preserving areal weighting, or through pycnophylactic interpolation. I reused a pre-existing R package and added several extensions, which include allowing the input of the aggregated data to be an ESRI Shapefile that associates polygonal regions to particular counts, or adding the support for a sampling procedure that spreads the sampled locations uniformly over the geographic space.

- **A detailed evaluation of the procedure**: I conducted experiments using different regression models to combine the different sources of ancillary information, with particular emphasis on state-of-the-art methods that estimate different weights to each of the ancillary variables over the geographic space (i.e., geographically weighted regression), and on methods that work better for data that does not meet the general assumptions of standard linear regression, like robust regression. Apart from few exceptions, the differences between the different types of regression algorithms were always very small, and the model that achieved the best results in the experiments varied in each case study.

- **Experiments with social media data**: The density of Flickr photos was used to aid in the spatial disaggregation approach, producing notably better results in a study concerning with the disaggregation of tourism statistics in the territory of Belgium. Due to the high level of Flickr activity in that region, a generalized additive regression model could effectively take advantage of the extra information provided by Flickr, and use the correlation of this data with the indicator under study to produce lower errors in the disaggregation.

5.2 Future Work

For future work, it would be interesting to continue improving the spatial disaggregation methodology. For instance, in the experiments that were reported in this dissertation, I have already compared different types of regression models, although other approaches could also be interesting to test. Other robust regression methods, such as least trimmed squares (Rousseeuw and Leroy, 1987), different Huber M-estimators or M-quantile models (Andersen, 2008; Chambers and Tzavidis, 2006; Salvati et al., 2012; Schmid and Münnich, 2014; Giusti et al., 2014), can for instance provide estimates with a superior quality, in the presence of outliers or when the

classical assumptions of linear regression are not met. Methods based on multi-layered neural networks, for instance involving convolutions over the raster data, could also be interesting to test (Maggiori et al., 2016; Nogueira et al., 2016). In fact, deep convolutional networks are nowadays the state-of-the-art approach in a related problem to that of spatial downscaling, namely in the context of image super-resolution (Dong et al., 2016; Kim et al., 2016).

Another possible direction for future research is the enrichment of the proposed approach, with the association of variance values to the disaggregation results, resulting in the production of fine-resolution estimates together with associated measures of uncertainty (Whitworth et al., 2016; Nagle et al., 2014). For example, a bootstrapping approach based on running the disseveration procedure multiple times, with random samples from initial estimates (i.e., random samples taken from the raster produced in Step 2 of the methodology outlined in Section 3.1), can be used to produce a raster with uncertainty values (i.e., the variance over the different tests), associated to the downscaled estimations.

It is my belief that the proposed approach is an important contribution to the literature, by combining the ideas of dasymetric mapping and pycnophylactic interpolation to disaggregate/downscale different types of indicators. However, recent studies in the area have also proposed other types of downscaling methods, for instance based on fractal analysis and interpolation (Vega, 2012; Kim and Barros, 2002; Xu et al., 2015; Sémécurbe et al., 2016). The combination of these different approaches, with their respective strengths and flaws, can perhaps result in lower error metrics in the disaggregation procedure. A simple way to include the aforementioned methods in the disseveration algorithm is to compute distinct ancillary rasters with their respective estimates, and use them as ancillary variables in the regression procedure.

Yet another idea for future work concerns with the usage of other types of ancillary data, in addition to the ones reported in this dissertation. For instance, the Worldgrids repository[1] contains several original or reformatted/reprojected global environmental layers, divided into themes like administrative and socio-economic data, land cover and land use, natural hazards, urbanization, and lights at night images, among others. A particular raster available from this repository that can be useful in the context of spatial disaggregation applications contains the distance of each raster cell to the nearest coastline, in kilometers. The map is derived from the global dataset named World Vector Shoreline[2], a digital data file containing the shorelines, international boundaries, and country names of the world. In addition, a raster showing the

[1] `http://worldgrids.org`
[2] `http://http://gcmd.nasa.gov/records/GCMD_WVS_DMA_NIMA.html`

travel time to major cities, originally outlined by Nelson (2008), is also made available[3]. The production of this last map involved a variety of global accessibility values, together with population density, with the main purpose of measuring the concentration of economic activity in a way that allows for a clear distinction of rural versus urban regions. It is my belief that this singular dataset is also highly correlated with different types of socio-economic indicators.

Another possible source of interesting ancillary data is the London Datastore[4], which is a web portal that contains statistical data on topics like crime levels, air pollution, and housing conditions, among others. Previous work (Blanchflower and Oswald, 2000; Jivraj and Nazroo, 2014) found that people with a higher socioeconomic status experience higher level of wellness, and that socioeconomic factors seem to be highly explanatory of well-being, although they do not fully explain it. Taking this into account, and considering that urban form also plays an important role in defining the well-being of city residents, and perhaps also other socio-economic characteristics, a dataset with measures of the proportion of historic properties in a given city area would be interesting to be used as ancillary information in disaggregation procedures. These values could, for instance, be computed through a simple formula that divides the number of historic buildings in a given area, extracted from the London Datastore, by the total value of the case study region. Using the same intuition, geo-tagged social media data obtained from Foursquare[5], a social media platform where the users share their whereabouts by checking-in into places/amenities and get recommendations on where to go based on their check-in history, could also be helpful. Several measures that can be obtained from these sources were found to be particularly interesting in studies similar to the one reported in this dissertation (Venerandi et al., 2015), such as the percentage of factories, the percentage of golf courses, or the number of wine shops. In cities with high user penetration, services like FourSquare constitute an accurate land use dataset. However, as with any source of volunteered geographic information, this approach may have geographic biases, and check-ins are mostly concentrated in city centers.

Finally, for future work, there are also plans to perform spatial analysis on historical census data, a task that involves additional challenges related to boundary changes, over time, of the administrative units. This problem is often circumvented through the aggregation of the data to a coarser level, which emphasises the modifiable areal unit problem, i.e., the relationships between census variables and other indicators vary from one source zone to another. Gregory (2002) has, for instance, reported on case studies over the Great Britain Historical Geograph-

[3] http://worldgrids.org/doku.php?id=wiki:urbanization_and_lights_at_night_images
[4] http://data.london.gov.uk
[5] http://foursquare.com

ical Information System[6], that contains the different boundaries and a large statistical census database for the United Kingdom, from the mid-nineteen century to 1994. He reported on experiments concerned with disaggregating the data collected at a coarse level to the finest level of spatial detail possible, using seminal methods from the spatial analysis literature. At this fine level, Gregory (2002) managed to compare the census disaggregated estimates from different years, avoiding the modifiable areal unit problem. Similar research studies, based on the same idea but using the more advanced disseveration procedure, can now easily be carried out.

[6] http://www.port.ac.uk/research/gbhgis

Bibliography

Andersen, R. (2008). *Modern Methods for Robust Regression.* Number 152 in Quantitative Applications in the Social Sciences. Sage Publications.

Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2012). *Applied Spatial Data Analysis with R.* Springer.

Blanchflower, D. G. and Oswald, A. J. (2000). Well-Being over time in Britain and the USA. Technical Report 7487, National Bureau of Economic Research.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1).

Briggs, D. J., Gulliver, J., Fecht, D., and Vienneau, D. M. (2007). Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sensing of Environment*, 108(4).

Burrough, P. A. and McDonnell, R. A. (1998). *Principles of geographical information systems.* Spatial Information Systems and Geostatistics. Oxford University Press.

Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3).

Chakir, R. (2009). Spatial downscaling of agricultural land-use data: an econometric approach using cross entropy. *Land Economics*, 85(2).

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93(2).

Cohen, W. W., Ravikumar, P. D., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI Workshop on Information Integration on the Web.*

Cromley, R. G., Hanink, D. M., and Bentley, G. C. (2011). A quantile regression approach to areal interpolation. *Annals of the Association of American Geographers*, 102(4).

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1).

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45).

Doll, C. N. H., Muller, J.-P., and Elvidge, C. (2000). Night-time imagery as a tool for global mapping of socio-economic parameters and greenhouse gas emissions. *Ambio*, 29(3).

Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2).

Douglass, R., Meyer, D., Ram, M., Rideout, D., and Song, D. (2015). High resolution population estimates from telecommunications data. *European Physical Journal Data Science*, 4(1).

Egerstedt, M. and Martin, C. (2009). *Control Theoretic Splines: Optimal Control, Statistics, and Path Planning*. Princeton Series in Applied Mathematics. Princeton University Press.

Eicher, C. L. and Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28(2).

Elvidge, C., Erwin, E., Baugh, K., Ziskin, D., Tuttle, B., Ghosh, T., and Sutton, P. (2009). Overview of DMSP nightime lights and future possibilities. In *Proceedings of the Joint Urban Remote Sensing Event*.

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., Davis, E. R., and Davis, C. (1997). Relation between satellite observed visible to near infrared emissions, population, and energy consumption. *International Journal of Remote Sensing*, 18(1).

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

Fotheringham, A. S., Charlton, M. E., and Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, 30(11).

Gallego, F. J. (2010). A population density grid of the European Union. *Population and Environment*, 31(6).

Giusti, C., Tzavidis, N., Pratesi, M., and Salvati, N. (2014). Resistance to outliers of M-quantile and robust random effects small area models. *Communications in Statistics-Simulation and Computation*, 43(3).

Goodchild, M. F., Anselin, L., and Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A*, 25(3).

Goodchild, M. F. and Lam, N. S.-N. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo-processing*, 1(1).

Gregory, I. (2002). The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, 26(4).

Guo, L., Ma, Z., and Zhang, L. (2008). Comparison of bandwidth selection in application of geographically weighted regression: a case study. *Canadian Journal of Forest Research*, 38(9).

Harris, P., Fotheringham, A. S., Crespo, R., and Charlton, M. (2010). The use of geographically weighted regression for spatial prediction: An evaluation of models using simulated data sets. *Mathematical Geosciences*, 42(6).

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall.

Hauff, C., Thomee, B., and Trevisiol, M. (2013). Working notes for the placing task at MediaEval 2013. In *Proceedings of the Workshop of the MediaEval Benchmarking Initiative for Multimedia Evaluation*.

Hawley, K. and Moellering, H. (2005). A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science*, 32(4).

Heymann, Y., S. C. C. G. and Bossard, M. (1994). CORINE land cover technical guide. Technical Report EUR12585, Office for Official Publications of the European Communities.

Huber, P., Wiley, J., and InterScience, W. (1981). *Robust statistics*. Wiley New York.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5).

Hwang, J.-N., Lay, S.-R., and Lippman, A. (1994). Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing*, 42(10).

Jiang, S., Alves, A., Rodrigues, F., Jr., J. F., and Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53(1).

Jivraj, S. and Nazroo, J. (2014). Determinants of socioeconomic inequalities in subjective well-being in later life: a cross-country comparison in England and the USA. *Quality of Life Research*, 23(9).

Kim, G. and Barros, A. P. (2002). Downscaling of remotely sensed soil moisture with a modified fractal interpolation method using contraction mapping and ancillary data. *Remote Sensing of Environment*, 83(3).

Kim, H. and Yao, X. (2010). Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method. *International Journal of Remote Sensing*, 31(21).

Kim, J., Kwon Lee, J., and Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5).

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, 84(408).

Li, D., Zhao, X., and Li, X. (2016). Remote sensing of human beings - a perspective from nighttime light. *Geo-spatial Information Science*, 19(1).

Lin, J. and Cromley, R. G. (2015a). Evaluating geo-located Twitter data as a control layer for areal interpolation of population. *Applied Geography*, 58(1).

Lin, J. and Cromley, R. G. (2015b). A local polycategorical approach to areal interpolation. *Computers, Environment and Urban Systems*, 54(1).

Lin, J., Cromley, R. G., and Zhang, C. (2011). Using geographically weighted regression to solve the areal interpolation problem. *Annals of GIS*, 17(1).

Longley, P. A. and Adnan, M. (2015). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, 30(2).

Longley, P. A., Adnan, M., and Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, 47(2).

Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P. (2016). Fully convolutional neural networks for remote sensing image classification. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*.

Malone, B. P., McBratney, A. B., Minasny, B., and Wheeler, I. (2012). A general method for downscaling Earth resource information. *Computers and Geosciences*, 41(2).

Matisziw, T., Grubesic, T., and Wei, H. (2008). Downscaling spatial structure for the analysis of epidemiological data. *Computers, Environment and Urban Systems*, 32(1).

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55(1).

Mennis, J. and Hultgren, T. (2006). Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. *Cartography and Geographic Information Science*, 33(3).

Nagle, N. N., Buttenfield, B. P., Leyk, S., and Spielman, S. (2014). Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104(1).

Nelson, A. (2008). Travel time to major cities: A global map of Accessibility. Global Environment Monitoring Unit - Joint Research Centre of the European Commission, Ispra Italy. Available at http://gem.jrc.ec.europa.eu/.

Nogueira, K., Penatti, O. A. B., and dos Santos, J. A. (2016). Towards better exploiting convolutional neural networks for remote sensing scene classification. *CoRR*, abs/1602.01517.

Nordhaus, W. D. (2003). Alternative approaches to spatial rescaling. Technical Report 1, Yale University, New Haven.

Nordhaus, W. D. (2006). Geography and macroeconomics: New data and new findings. *Proceedings of the National Academy of Sciences*, 103(10).

Patel, N. N., Stevens, F. R., Huang, Z., Gaughan, A. E., Elyazar, I., and Tatem, A. J. (2016). Improving large area population mapping using geotweet densities. *Transactions in GIS*, 15(1).

Perez-Verdin, G., Marquez-Linares, M. A., and Salmeron-Macias, M. (2014). Spatial heterogeneity of factors influencing forest fires size in northern Mexico. *Journal of Forestry Research*, 25(2).

Quinlan, R. J. (1992). Learning with continuous classes. In *Proceedings of the Australian Joint Conference On Artificial Intelligence*.

Reibel, M. and Bufalino, M. E. (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37(1).

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*. John Wiley & Sons, Inc.

Salvati, N., Tzavidis, N., Pratesi, M., and Chambers, R. (2012). Small area estimation via M-quantile geographically weighted regression. *Test*, 21(1).

Schmid, T. and Münnich, R. T. (2014). Spatial robust small area estimation. *Statistical Papers*, 55(3).

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. Wiley series in probability and mathematical statistics : Applied probability and statistics section. Wiley-Interscience.

Sémécurbe, F., Tannier, C., and Roux, S. G. (2016). Spatial distribution of human population in France: Exploring the modifiable areal unit problem using multifractal analysis. *Geographical Analysis*, 48(3).

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3).

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the ACM National Conference*.

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. (2015). High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific Data*, 2(1).

Sridharan, H. and Qiu, F. (2013). A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geographical Analysis*, 45(3).

Steiger, E., Westerholt, R., Resch, B., and Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, 54(1).

Stein, M. L. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE*, 10(2).

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367).

Tobler, W., Deichmann, U., Gottsegen, J., and Maloy, K. (1995). The global demography project. Technical Report 95-6, National Center for Geographic Information and Analysis, Santa Barbara.

Vanwambeke, S. O., Bennett, S. N., and Kapan, D. D. (2011). Spatially disaggregated disease transmission risk: land cover, land use and risk of dengue transmission on the island of Oahu. *Tropical Medicine International Health*, 16(2).

Vega, K. V. A. (2012). *Aplicación de la interpolación fractal en downscaling de imágenes satelitales NOAA-AVHRR de temperatura de superficie en terrenos de topografia compleja.* PhD thesis, Universidad de Chile.

Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S.* Springer.

Venerandi, A., Quattrone, G., Capra, L., Quercia, D., and Sáez-Trumper, D. (2015). Measuring urban deprivation from user generated content. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing.*

Whitworth, A., Carter, E., Ballas, D., and Moon, G. (2016). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. *Computers, Environment and Urban Systems.*

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1).

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers Inc.

Wu, S.-s., Qiu, X., and Wang, L. (2005). Population estimation methods in GIS and remote sensing: A review. *GIScience & Remote Sensing*, 42(1).

Xu, G., Xu, X., Liu, M., Sun, A. Y., and Wang, K. (2015). Spatial downscaling of TRMM precipitation product using a combined multifractal and regression approach: Demonstration for south China. *Water*, 7(6).

Zhang, C. and Qiu, F. (2011). A point-based intelligent approach to areal interpolation. *The Professional Geographer*, 63(2).