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

João Miguel Cordeiro Monteiro

Abstract

Although statistical information on socio-economic activities is widely available, 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 into a set of target zones, with different geometry and with a higher general level of spatial resolution. This paper presents a 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, OpenStreetMap road density, or geo-referenced data extracted from Flickr, to disaggregate different types of socio-economic indicators.

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. However, using aggregated data usually masks important local hotspots, and overall tends to smooth out spatial variations in impact. For this reason, researchers often need to disaggregate source data, in order to provide more localized estimates. In the context of spatial analysis, 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 used to convert data originally available for a set of source zones into a set of target zones, that have a different geometry and a higher general level of spatial resolution. They have in common what Tobler (1970) termed the pycnophylactic, or mass-preserving, property, in that the estimates are conditioned to sum to the original quantities in the source zones. The term spatial disaggregation is in fact usually employed in the context of additive variables (e.g., population counts), whereas spatial downscaling is more general, being frequently used in the context of procedures that focus on non-additive variables (e.g., different types of environmental or geophysical properties, such as temperature).

In this paper, I report on experiments with an hybrid spatial disaggregation technique that combines the ideas of dasymetric mapping and pycnophylactic interpolation, using population density, nighttime satellite imagery, land coverage information, and OpenStreetMap road density information, together with ancillary data extracted from popular location-based services and social media sources, like Flickr, to disaggregate different types of socio-economic indicators. Apart from few exceptions (e.g., seminal work on the area by Goodchild et al. (1993) that considered variables such as employment and income), most previous studies concerning with spatial disaggregation/downscaling have focused either on population density or on geophysical/environmental variables. I nonetheless believe that previously developed procedures, which had these traditional applications in mind, can also be equally useful in the domain of socio-economic
indicators. The spatial disaggregation technique discussed in this paper resulted in the production of raster datasets with different types of socio-economic indicators. A raster dataset is a type of mosaic that divides a surface into uniform cells, making it easier to examine socio-economic data through different partitions of space, or in terms of their relation to particular geophysical characteristics (e.g., proximity to regions with specific land coverage types, or relationship towards terrain elevation).

The rest of this paper is organized as follows: Section 2 presents fundamental concepts and important related work. Section 3 describes the considered hybrid spatial disaggregation approach. Section 4 details the case studies concerning with socio-economic indicators. Finally, Section 5 presents the main conclusions, and highlights possible directions for future work.

2 Fundamental Concepts and Related Work

The most basic method for spatial disaggregation is mass-preserving areal weighting, in which a homogeneous distribution of the data throughout each source zone is assumed (Goodchild and Lam, 1980). Mass-preserving areal interpolation redistributes the aggregated data with basis on the proportion of each source zone that overlaps with the target zone, according to the following equation:

\[
P_t = \sum_{\{s:s \cap t \neq \emptyset\}} \left( P_s \times \frac{A_{ts}}{A_s} \right)
\]

In the formula, the parameter \(P_t\) is the estimated count in a target zone \(t\), while \(P_s\) is the count in a source zone \(s\) that is to be disaggregated. The parameter \(A_s\) corresponds to the area of source zone \(s\), and \(A_{ts}\) corresponds to the area of target zone \(t\) overlapping with the source zone \(s\). While areal weighting disaggregation ensures that the total count from the source data remains unchanged, it is based on the often incorrect assumption that the phenomena of interest are evenly distributed across the source zones. Population is one such example where the assumption behind mass-preserving areal interpolation clearly does not hold, since 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.

Mask area weighting, also referred to as binary dasymetric mapping, is an improvement on simple area weighting in that it uses a mask to define where, within the target zone, the source data should be allocated (Eicher and Brewer, 2001). Each source unit is divided into two sub-regions (i.e., populated and unpopulated) and the source information is then allocated only to the populated areas. General dasymetric disaggregation is an improvement on mask areal weighting, in that two or more categories can be assigned weights for disaggregation (e.g., derived for individual land cover types to reflect population density). The weights of each of the categories for the source area represent the percentage of population (or another variable) that is likely to be contained within that category, per source area. The main challenge in dasymetric disaggregation thus involves devising an appropriate set of weights that can be applied to the ancillary data. Weights may, for instance, be defined using selective sampling, or by some form of regression analysis. The general dasymetric disaggregation method that corresponds to an extreme case where there is not a predefined number of classes corresponds to a proportional and weighted areal interpolation method, whose formula is shown below, where \(W_t\) is the weight assigned to the target zone \(t\), and where each \(W_t\) is chosen with basis on the external variable(s), 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 (i.e., in the case of disaggregation).

\[
P_t = \sum_{\{s:s \cap t \neq \emptyset\}} \left( P_s \times \frac{W_t \times A_{ts}}{\sum_{\{t':t' \cap s \neq \emptyset\}} W_{t'} \times A_{t'}} \right)
\]

The above method would disaggregate the source data under the assumption that target regions containing a higher value for the external variable will also correspond to regions having higher counts in the source data. This idea is explored on the present paper.
Recently, Malone et al. (2012) described a spatial downscaling algorithm based on a disseveration principle, also providing an easily extensible open-source R implementation that was used in the experiments reported on this paper (i.e., although the original method was proposed for spatial downscaling and not for spatial disaggregation, I extended the implementation originally outlined by Malone et al. in order to develop a novel spatial disaggregation method). The algorithm described by Malone has two phases, and it originally used generalized additive modeling to fit a non-linear relationship between a target variable (i.e., the indicator that we wish to model at a fine resolution, and for which we have data at a coarse resolution) and predictive covariates (i.e., data for other variables, available at a fine resolution). In an initialization phase, the authors perform a coarse grid to fine grid re-sample (i.e., through a nearest neighbor re-sampling approach), followed by sampling and initial model fit. In an iteration phase, adjustments are made to the predictions iteratively, trying to ensure that the coarse grid map is linearly related to the fine grid predictions (i.e., there is a mass balance property to be attained). Iterations proceed until a stopping criterion is met, based on a maximum number of iterations (e.g., 100 iterations) or on a threshold of 0.001 over the change in the estimated error rate over three consecutive iterations.

Although dasymetric-mapping methods are preferable to other methods that do not use ancillary data, having been show to achieve a superior performance (Wu et al., 2005), they also have several methodological and cartographic shortcomings, one of them being that the estimated density values often change abruptly. Rather than performing the disaggregation into zones, several methods instead entail at creating continuous surfaces depicting the disaggregation of the data. For instance Tobler (1979) proposed one such pycnophylactic spatial disaggregation method, which is an extension of simple areal weighting that assumes a degree of spatial auto-correlation of the variable being interpolated. Tobler's method starts by applying the mass-preserving areal weighting procedure described previously, using a grid to define the target zones. Then, the values for the grid cells $P_t$ are smoothed, by replacing them with the average of their four neighbors. The predicted values in each source zone are then compared with the actual values, and adjusted to meet the pycnophylactic condition of mass-preservation, continuing until there is either no significant difference between predicted values and actual values within the source zones, or until there have been no significant changes of cell values from the previous iteration. The interpolated surface produced by the pycnophylactic method by Tobler (1979) is smooth, with relatively small changes in attribute values at target region boundaries.

Authors such as Kim and Yao (2010) have also developed hybrid approaches for the spatial disaggregation of population data that combine dasymetric mapping and pycnophylactic interpolation, making use of ancillary information that sheds light on the spatial structure of population distribution, and also adopting the conceptual assumption that population density varies smoothly, instead of uniformly, in space.

### 3 An Hybrid Method for Spatial Disaggregation

Both the dasymetric mapping and the pycnophylactic interpolation methods have solid theoretical foundations, as well as strong empirical support in population-estimation research. Each of these methods has its own strengths, but also suffers obvious shortcomings. In this paper, I present 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). Kim and Yao (2010) proposed a method that starts with binary dasymetric disaggregation leveraging land coverage data, for producing initial estimates that are latter refined through pycnophylactic interpolation. In my work, I propose to leverage pycnophylactic interpolation for producing initial estimates, that are then adjusted through a methodology based on disseveration (Malone et al., 2012). The general procedure is detailed next, trough an enumeration of all the individual steps that are involved:

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, corresponding to an initial estimate for the
disaggregated values. This raster, referred to as $T^p$, will contain smooth values resulting from a pycnophylactic interpolation procedure (Tobler, 1979);

3. Overlay four rasters $P^1$, $P^2$, $P^3$ and $P^4$ on the study region 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;

4. Overlay two other rasters $P^5$ and $P^6$ over the study region with ancillary information derived from the rasters in the previous step. Specifically, these two 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 $P^3$, whereas raster $P^6$ is derived from raster $P^4$;

5. Overlay another raster $T^d$ on the study region, storing the estimates produced by a simple spatial disaggregation procedure based on dasymetric mapping. For producing these estimates, the total value is weighted, for each source zone in the original thematic map, according to 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;

6. Create yet another raster $P^7$ with the distribution of social media usage under the geographic region. 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;

7. Collect a sample of cells in the fine resolution grid, in order to fit the regression models, either using regular (i.e., systematically aligned) sampling, or clustered sampling (i.e., the same number of samples are collected from groups of points assumed to have different characteristics);

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 rasters $P^1$, $P^2$, $P^3$, $P^4$, $P^5$, $P^6$ and $T^d$, as predictive covariates. The regression algorithm used in the disseveration procedure is fit using the data sample from the previous step;

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;

10. Steps 7 to 9 are repeated, by running iteratively 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.

The above procedure was implemented through the programming language of the R\footnote{http://www.r-project.org} project for statistical computing. I specifically relied on the R packages named pycno\footnote{http://cran.r-project.org/web/packages/pycno/index.html} and dissever\footnote{http://github.com/pierreroudier/dissever}, 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. 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 covariates. In the experiments, I specifically 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). 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 based on a combination of multiple decision trees, which will attempt to approximate the linear relation with an irregular step function). For more complex relationships between the target values and the covariates, non-linear models can perhaps offer a better performance.

In standard linear regression, a global linear least-squares fit is computed for a set of predictor variables (i.e., the covariates) to predict 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. In generalized additive models, each of the predictor variables are instead connected to the dependent variable via a link function, instead of estimating single parameters, effectively allowing for some degree of non-linearity. The cubist approach, originally proposed by Quinlan (1992), combines decision trees with linear regression models. The leaf nodes in these decision trees contain linear regression models that are smoothed by taking into account the predictions from the linear models in the previous nodes, recursively up the tree. Robust regression approaches are designed to circumvent some of the limitations that traditional methods have, namely in handling 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). The technique that was used tries to minimize the sum of each residual contribution, by multiplying each of the residual value by an assigned weight. Finally, when considering geographically weighted regression, the focus is on the identification of several local relationships between variables, i.e., locally weighted regression coefficients concerning different geographic zones, taking into account the phenomena of spatial nonstationarity (i.e., different intrinsically disparities of contextual issues).

The ancillary information was acquired from the following datasets:

- Population statistics came from the Gridded Population of the World (GPW4), a well-known dataset depicting the distribution of human population across the globe, providing globally consistent and disaggregated population information. In the experiments, the GPW dataset considered was the count data, projected to the year of 2010 and with the resolution of 30 arc-seconds per cell;
- Regarding nighttime light emissions, the experiments used the publicly available VIIRS Nighttime Lights-2012 dataset5, maintained by the Earth Observation Group of the NOAA National Geophysical Data Center. 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. 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 used the standard Corine Land Cover (CLC) data product6, which is based on satellite images as the primary information source. I specifically used data for the year of 2012, on a 250 × 250 meters resolution. The 44 different classes of the 3-level Corine nomenclature that are considered in the original product were converted into a real value in the range [0, 1], which encodes how developed is the territory corresponding to a given cell, making it easier to explore land coverage within different types of regression modeling methods. Besides the raster encoding land development, the CLC dataset was also used to produce a second raster with derived information, encoding the distance towards the nearest water body;
- Regarding OpenStreetMap information, I used the data from a study put forward by Martin Raifer in 2015, that outlines the most densely mapped regions in OpenStreetMap7 (i.e., the number of street segments per cell). From different shapefiles containing OpenStreetMap road network information8,
an additional dataset with derived information was also produced, encoding the distance from a
given position (i.e., from a given cell) towards the nearest road or street segment. In both cases, the
considered resolution was of 30 arc-seconds per cell.

- Finally, the social media data that were considered in some of the experiments were originally made
available in the context of the location estimation subtask from the MediaEval Placing Task\(^9\), which
concerned with the development of data-driven methods to estimate the latitude/longitude coordinates
at which a photograph was taken. Specifically, I relied on the publicly available information
from the 2013\(^10\) and 2015\(^11\) editions of the placing task, to compute a raster containing the density
of the Flickr photos from both MediaEval datasets. The coordinates of all the images were used to
produce a map with their counts per each raster cell, and then a kernel density estimation function
was applied to the previous raster to smooth the count data.

4 Experimental evaluation

In my case studies, I used socio-economic data pertaining to the Portuguese, Belgian, and French territories,
and its administrative units, publicly available from the Portuguese National Institute of Statistics, the
annual report of the observatory for tourism in Brussels, or from the Eurostat portal. I have specifically
used the following variables in the case studies:

- Number of female residents in Portugal, according to the national census in 2011;
- Number of live births in Portugal, by place of residence of the mother, in 2011;
- Number of deaths in Portugal, according to the national directorate-general of health, in 2011;
- Number of foreign residents in Portugal, according to the national census in 2011;
- Number of buildings in Portugal, according to the national census in 2011;
- Number of buildings with at least two floors in Portugal, according to the national census in 2011;
- Resident population employed in the agriculture, animal production, hunting, forest, and fishery
  sectors in Portugal, according to the national census in 2011;
- Employed resident population in Portugal, according to the national census in 2011;
- Number of crimes registered by the police forces in Portugal, in 2011;
- Number of hotel visitors (i.e., number of guests in hotel establishments) in Portugal, according to
  the national tourism authority, in 2011;
- Number of overnights, according to the annual report of the observatory for tourism in Brussels, in
  2012;
- Number of nights spent at tourist accommodation establishments for the territory of France, accord-
  ing to the Eurostat regional tourism statistics, in 2012.

For testing the spatial disaggregation procedure in an initial case study, the 4260 Portuguese civil
parishes were first used as the source units, producing raster datasets with a resolution of 30 arc-seconds
per cell. Exceptions to this procedure are the variables corresponding to the number of crimes and the

\(^9\)http://www.multimediaeval.org
\(^10\)http://www.st.ewi.tudelft.nl/~hauff/placingTask2013Data.html
\(^11\)http://repository.tudelft.nl/islandora/object/uuid:ec44e1d0-1228-463a-a6c2-e693b8091bc1
Figure 1: Aggregated data for the different socio-economic indicators.

number of hotel visitors in Portugal, for which one only had access to data aggregated at the level of municipalities.

Figure 1 presents a grid with multiple choropleth maps, illustrating the aggregated information at the level of civil parishes, for the considered socio-economic indicators and for the administrative units in Continental Portugal. 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 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 were no information was available. All the maps from Figure 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 2 presents a similar grid to the one that is shown in Figure 1, but in this case illustrating the results that were obtained through the proposed spatial disaggregation procedure, using as source zones the highest possible resolutions 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 algorithm. The maps from Figure 2 have a resolution of 30 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 3 details one of the variables from Figures 1 and 2, specifically the one corresponding to 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
Figure 2: Disaggregation results for the different socio-economic indicators.

Portuguese territory, (iii) a raster showing the disaggregated number of foreign residents, as obtained with the simpler method that only used population data in the disaggregation 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 data. 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 number of foreigners. Also, the resulting map is smoother than the one that would be produced by the proportional and weighted areal interpolation procedure.

It should be noted that spatial disaggregation is never an error-free process, and errors introduced during disaggregation can be propagated to the subsequent analysis steps. The typical accuracy assessment strategy is to aggregate the target zone estimates to either the source or some intermediary zones, and
then compare 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 are as follows.

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

(3)

\[
MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}
\]

(4)

In Equations (3) and (4), \(\hat{y}_i\) corresponds to a predicted value, \(y_i\) corresponds to a true value, and \(n\) is the number of predictions. This article also reports results in terms of the Normalized Root Mean Square Error (NRMSE) and the Normalized Mean Absolute Error (NMAE) to facilitate the comparison across variables, in which the values of the RMSE and MAE are divided by the amplitude of the true values.

To get some idea on the errors involved in the proposed spatial disaggregation procedure, I experimented with the disaggregation of data originally reported at the level of large territorial divisions (i.e., at the level of municipalities) to the raster level, latter aggregating the estimates to the level of civil parishes (i.e., taking the sum of the values from all raster cells associated to each civil parish) and comparing the aggregated estimates against the values that were originally available for the 4260 civil parishes.

Tables 1 and 2 show the obtained results. Table 1 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 (i.e., raster \(T^d\) in the enumeration given in Section 3). Table 2 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 based on the cubist method. The results for the NRMSE and NMAE metrics are reported with a multiplication factor of \(10^{-2}\). Values shown in bold correspond to the best results that were achieved for each variable (i.e., values in bold shown in Table 1 correspond to cases in which the hybrid disaggregation method, based on disseveration, could not outperform one of the baselines, namely the one based on weighted areal disaggregation leveraging population).

The results from Tables 1 and 2 show that the proposed hybrid method indeed outperforms the baselines corresponding to areal interpolation, pycnophylactic interpolation, or weighted areal interpolation, at a municipality level. However, in some error metrics and particularly for indicators that have a strong

<table>
<thead>
<tr>
<th>Areal Interpolation</th>
<th>Pycnophylactic Interpolation</th>
<th>Weighted Interpolation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Female residents</td>
<td>2220.60</td>
<td>876.94</td>
</tr>
<tr>
<td>Live births</td>
<td>47.21</td>
<td>17.40</td>
</tr>
<tr>
<td>Deaths</td>
<td>35.98</td>
<td>15.57</td>
</tr>
<tr>
<td>Foreign residents</td>
<td>322.70</td>
<td>82.41</td>
</tr>
<tr>
<td>Buildings</td>
<td>786.00</td>
<td>412.58</td>
</tr>
<tr>
<td>Primary sector</td>
<td>41.83</td>
<td>16.97</td>
</tr>
<tr>
<td>Employees</td>
<td>1871.29</td>
<td>724.62</td>
</tr>
</tbody>
</table>

Table 1: Disaggregation errors measured for different socio-economic variables, using baseline methods and with the aggregated data collected originally at the level of municipalities.

<table>
<thead>
<tr>
<th>Linear Models</th>
<th>Generalized Additive Models</th>
<th>Cubist</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Female residents</td>
<td>580.04</td>
<td>188.44</td>
</tr>
<tr>
<td>Live births</td>
<td>14.99</td>
<td>5.97</td>
</tr>
<tr>
<td>Deaths</td>
<td>16.55</td>
<td>7.40</td>
</tr>
<tr>
<td>Foreign residents</td>
<td>133.31</td>
<td>34.78</td>
</tr>
<tr>
<td>Tall buildings</td>
<td>311.83</td>
<td>154.74</td>
</tr>
<tr>
<td>Prim. sect. workers</td>
<td>37.37</td>
<td>15.33</td>
</tr>
<tr>
<td>Employed pop.</td>
<td>503.83</td>
<td>169.80</td>
</tr>
</tbody>
</table>

Table 2: 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.
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. When the indicator to disaggregate is strongly correlated with population counts (e.g., for variables such as live births, or employed population), the methods that produced lower disaggregation errors used regression analysis based on standard linear regression or generalized additive models, instead of 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 the cases, the simple method based on weighted areal disaggregation, leveraging population data for the weights, indeed corresponded to a very strong baseline. Nonetheless, for almost all indicators, the usage of additional ancillary information can indeed lead to improvements, sometimes considerable ones.

Two other case studies concerning the territories of Belgium and France were considered, using the aggregated data collected from the referenced tourism statistics portals. The error results were compared when using the same regression models tested in the Portuguese case study, as well as a robust linear model and geographically weighted regression. In the disaggregation of the number of overnights in the Belgian territory, the divisions considered were three provinces, concerning the region of Brussels and its outskirts, that are sub-divided into 59 municipalities. On the other hand, the divisions considered in the experiments for the French territory were 8 French NUTS I divisions, that are sub-divided into 22 NUTS II divisions. Two experiments were conducted in each case study, one leveraging only the baseline ancillary information, and another using also the Flickr dataset. Apart from few exceptions, the error differences obtained when using the different types of regression algorithms, or when introducing Flickr as ancillary information, were always very small, and the model that achieved the best results in the experiments varied in each case study. However, when using the generalized additive model to leverage the density of Flickr photos in the disaggregation of the number of overnights in the Belgian territory, notably better results were produced. Due to the high Flickr activity in that region, this model could effectively take advantage of the extra information.

5 Conclusions and Future Work

In this paper, 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 OpenStreetMap, and social media data extracted from Flickr, as ancillary data to disaggregate different types of socio-economic indicators to a raster-grid level. The proposed disaggregation technique was applied in different case studies relative to the Portuguese, Belgian, and French territories, resulting in the production of fine-gridded rasters. The spatial disaggregation methodology was discussed, as well as the quality of the obtained results.

For future work, it would be interesting to continue improving the spatial disaggregation methodology. For instance, other regression methods based on multi-layered neural networks, involving convolutions over the raster data, could be interesting to test (Nogueira et al., 2016). Estimates for the variance associated with the disaggregation results could also be computed, resulting in the production of fine-resolution estimates together with associated measures of uncertainty (Whitworth et al., 2016). Another idea for future work concerns with the usage of other types of ancillary data, like the reformatted/reprojected global environmental layers available in the Worldgrids repository12. A particular raster available from this repository contains the distance of each raster cell to the nearest coastline, in kilometers. Another possible source of interesting ancillary data is the London datastore13, which is a web portal that contains statistical data on topics like crime levels, air pollution, and housing conditions, among others. Some of the information reported in this portal could effectively be used to aid in the disaggregation procedure.

12http://worldgrids.org
13http://data.london.gov.uk
References


