DISAGGREGATION OF REGIONAL POPULATION DATA FOR RESIDENTIAL HOT WATER DEMAND ASSESSMENT

Syed Monjur Murshed
European Institute for Energy Research (EIFER)
University of Karlsruhe,
Emmy-Noether Str. 11, 76131 Karlsruhe, Germany
murshed@eifer.uni-karlsruhe.de

Abstract

Population data is mostly available at different administrative units e.g. regional, districts and municipalities. But in many socio-economic fields of research, availability of such dataset at finer units is often very essential to carry out advanced analysis and improved decision making. The assessment of hot water demand in the residential sector is directly dependent on the number of inhabitants living in the households. Using the publicly accessible statistical data, it is possible to determine average hot water demand at certain administrative units. But such analysis only gives generalised overview; therefore, for an adequate decision, data needs to be further disaggregated. Estimation of the number of population living at finer land use unit would assist in more precise calculation of such demand. On the other hand, it might be easy to quickly assess the demand at a smaller area but it becomes even more complicated for a larger region. Therefore, a systematic method has been developed in this study.

The aim of the paper is to disaggregate regional population data, available at municipality level into land use units to ascertain the residential hot water demand. This study is carried out within the project “Renewable based Neighbourhood Heating of Baden-Württemberg (Green Heat)” at the “European Institute for Energy Research (EIFER)” with a view to assess the cost effectiveness of renewable based heating systems in the residential sector in the federal state of Baden-Württemberg in Germany.

The overall methodology of this study is implemented through different statistical and geographical analysis. At first, an extensive review on diasaggregation of population estimation methods in GIS and remote sensing is carried out. Afterwards, various statistical and geo data are prepared to feed into the regression models. In the second step, a regression based dasymetric mapping of areal interpolation is applied to the high resolution (25m × 25m) Infoterra LaND25 dataset to disaggregate population data into individual land use units. Finally, considering energy related parameters, hot water demand in each land use unit is calculated and afterwards visualised in maps.

The result is validated by comparing similar statistical data available at larger administrative units as well as other related studies. It is found out that with an accuracy usually comprised of between ±25%, the disaggregation of the population showed very good results in compared to the results of the state of the art. Regarding hot water

demand, the result of this study represents less than 15% of the total residential demand compared to the statistics. So the disparity of the results remains quite limited.

The outcome of the assessment of hot water demand within the residential land use units would help to optimise heating systems and assist the policy makers and the energy companies to know the potential of hot water market for Baden-Württemberg and thus facilitate in developing sustainable energy territories.

1. Introduction

Detailed socio-economic data are not generally available or are purposely aggregated to avoid problems of privacy. Under these circumstances, data have to be derived from larger spatial administrative zones where aggregated data are available (Huang and Ottens, 2007). One important type of such data is census data which offers a wide range of socio-economic information, but is aggregated within arbitrary enumeration districts. In many modelling situations, this aggregated data need to be disaggregated at other finer scale or are re-aggregated at coarser scale in order to derive added information.

The aim of this paper is to disaggregate census population data from larger administrative districts which is represented as municipality, into smaller units which is characterized as residential land use (LU) unit. Afterwards, residential hot water demands in these LU units are calculated. Therefore, a systematic statistical and spatial analysis techniques is applied at the residential LU classes of Infoterra LaND25 data (Infoterra, 2007), at a resolution of 25m × 25m, in each municipality in the federal state of Baden-Württemberg (BW). Geographic Information Systems (GIS) and statistical methods of SPSS software packages are exploited to carry out the analysis, thus providing a detailed representation of the hot water demand in BW.

This study is carried out within the project “Renewable based Neighbourhood Heating of Baden-Württemberg (Green Heat)” which was initiated by the “European Institute for Energy Research (EIFER)” with the financial support from the energy provider “Energie Baden-Württemberg AG (EnBW)” during 2007 and 2008. One of the main objectives of the project is to assess the cost effectiveness of renewable-based heating systems in the residential sector, by analysing the potential areas of space and hot water heating demand, quantifying the theoretical potentials of renewable energy resources. This paper mainly focuses on the calculation of hot water demand; the assessment of space heating demand is discussed in (Murshed et al., 2009).

In GIS and remote sensing literature, different methods of disaggregation of population data are elaborately discussed (Table 1). For example, (Wu et al., 2005) classified the methods into two main categories: areal interpolation and statistical modelling. Areal interpolation methods use census data as input and apply disaggregation techniques, with ancillary or without ancillary information, to obtain more refined population. On the other hand, statistical modelling infers relationship between population and other variables to estimate population of any particular region. Table 1 gives a detailed overview of different methods of disaggregation of population data.

<table>
<tr><th colspan="3">Disaggregation methods of population data</th></tr>
<tr><td colspan="2">a. Areal interpolation</td><td rowspan="2">b. Statistical modelling</td></tr>
<tr><td>With ancillary information</td><td>Without ancillary information</td></tr>
<tr><td rowspan="4">- Dasymetic mapping
i. Binary method
ii. Three class
iii. Regression</td><td>- Point-based Methods</td><td>- Correlation with Urban areas</td></tr>
<tr><td>- Pycnophylactic interpolation</td><td>- Correlation with LU</td></tr>
<tr><td>- Area-based Methods</td><td>- Correlation with Dwelling units</td></tr>
<tr><td>- Kernel-based interpolation</td><td>- Correlation with Image pixels characteristics</td></tr>
<tr><td>- Smoothing method</td><td></td><td>- Correlation with others physical and socio economic characteristics</td></tr>
</table>

Table 1. Diasaggregation methods of population data used in GIS and remote sensing

Areal interpolation method uses census population data as the input and applies interpolation techniques (Table 1) to obtain refined population data, usually for the purpose of transforming population data from source zone to the target zone. One advantage is that it considers the source zones as the operating units that have the volume preserving property. It means population data of the source zone would remain unchanged during the disaggregation process. Moreover, the interpolation method can be further improved by using ancillary information such as population, topography, land use, remote sensing image spectral and textural statistics, road networks, etc. This usually yields more accurate results than those without ancillary information, assuming the ancillary information reflects the spatial distribution of the variables being mapped (Eicher and Brewer, 2001), (Langford, 2006), (Mennis, 2003).

The simplest technique used for areal interpolation is known as areal weighting which can be adapted to distribute the data from statistical zones to LU units, by considering the weight of LU types (Huang and Ottens, 2007). When LU is used as ancillary information, areal interpolation is also referred to as dasymetric mapping. Among all the population estimation methods, the dasymetric mapping of areal interpolation is regarded as a more accurate approach, provided that the used ancillary information gives an authentic description of where people actually live. The dasymetric mapping approach is based on the following function:

$$P_i = \sum_j p_{ij} = \sum_j \left(A_{ij} \times D_j\right)$$

Where, P_i : total population of source zone *i*, p_{ij} : total population of LU *j* within source zone *i*, A_{ij} : total area of LU *j* within source zone *i*, D_j : average population density of LU *j*.

A_{ij} is obtained by GIS overlay operation, whereas D_j is generated using one of the dasymetric mapping methods. (Langford et al., 1991) mapped population density by assigning arbitrary population percentages to the Land use/land cover classes (LU/LC) with a binary method, (Mennis, 2003) introduced an empirical sampling method to assign meaningful population densities to the LU/LC classes (three class method). (Flowerdew and Green, 1989) proposed statistical regression analysis to estimate population densities. Among all these methods, regression analysis provides better results because of its objectivity in testing model accuracies through statistical significance test. (Wu et al., 2005), (Langford et al., 1991) and (Yuan et al., 1997) applied linear regression models to estimate the population density of the LU units.

This study proposed an improved regression analysis by categorising the statistical population data into four classes and then developing the regression models for each category. Moreover, the scaling technique is used to improve the results. These are discussed in Chapter 3. In Chapter 2 an introduction to the study area and the required data sets is given. The implementation of the methodology and calculation of hot water demand is illustrated next. Finally in Chapter 5, the results and conclusion are discussed by comparing with statistical data and by implying future research outlooks.

2. Description of study area and required dataset

The disaggregation method of population data is applied in the federal state of BW, Germany. It has 1,111 municipalities and 44 districts with a population of about 10.7 million residing in 35,751 km² area (DESTATIS, 2007). Baden-Württemberg is the 3rd largest state in Germany, but its high population induces a higher population density (300 inh/km²) than the average in Germany (231 inh/km²).

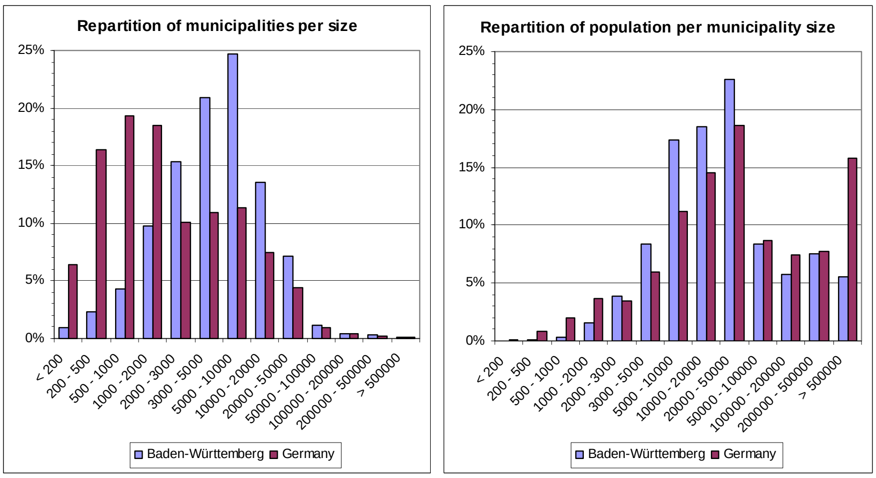

Figure 1. Distribution of municipality and number of inhabitants classified by municipality size in Baden-Württemberg (blue) and in Germany (purple), 2006, Source: (BMVBS, 2007)

Urban and residential structure of BW is quite different comparing the whole Germany (Figure 1). Medium-sized municipalities are widely distributed in BW than in the rest of Germany. While a larger share of the population live in this type of municipality (5,000 to 50,000 inhabitants) in BW: only 6% of the population lives in a city with more than 500,000 inhabitants (16% in Germany).

A wide variety of geo and statistical dataset is needed to carry out the analysis. Residential LU data is accumulated from the Infoterra LaND25 data which is available at a scale of 1:25,000 (Infoterra, 2007) as well as Corine land cover 2000 (CLC2000) data which is available at a scale of 1:100,000 (UBA et al., 2004). Population census data and geometric boundary data of each municipality are collected from the statistical office BW (StaLaBW, 2007) and INFAS (INFAS, 2001) respectively. Moreover, energy related parameters are referred from different literatures and statistics.

3. Proposed methodology to disaggregate population data

Considering the state of the art of disaggregation methods, aim of this study as well as the availability of data, regression based dasymetric mapping of areal interpolation is chosen to disaggregate population data from the municipality to the LU units.

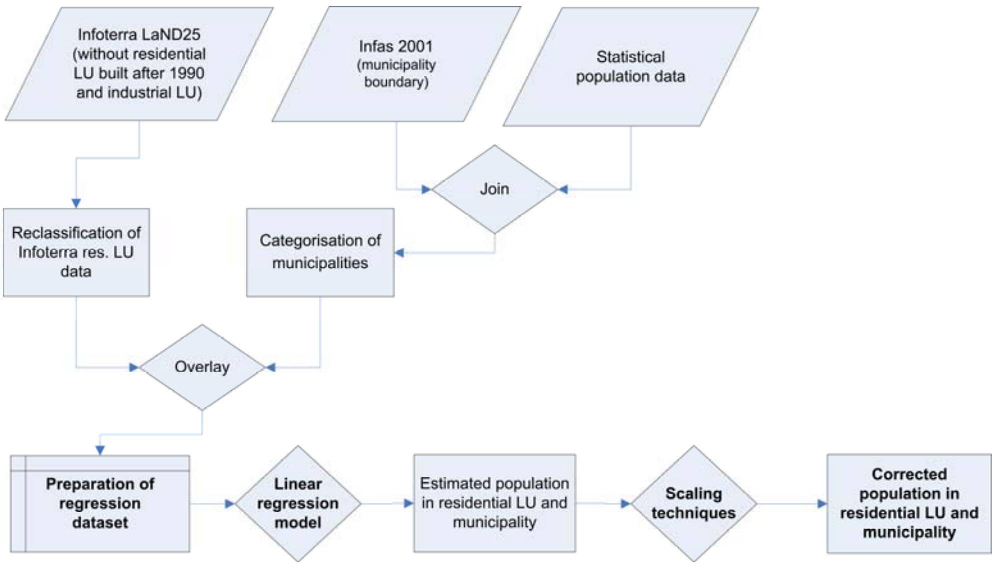

Figure 2. Methods of disaggregation of population data into residential LU units

The disaggregation methodology is divided into three main steps: (a) preparation of the regression dataset (b) development of regression model and (c) use of scaling techniques and estimation of population. The detailed methodology is illustrated in

Figure 2. The first step consists of reclassifying both the Infoterra LaND25 data into appropriate residential LU classes as well as the municipalities according to population number. Then the population data is overlayed with the reclassified Infoterra LaND25 data. The second step includes statistical regression analysis and modelling to estimate the population in the LU classes. In the final step, scaling technique is used to locally fit the estimated population. Each step is elaborately described in the following sections.

a) Preparation of the regression dataset

The municipalities in BW are reclassified according to the number of population. It will help to obtain improved regression results in different categories of municipalities and their corresponding LU classes. Therefore, all the 1,111 municipalities in BW are grouped into four different categories (Table 2) and then regression analyses are performed for each category.

Municipality category	**Number of inhabitants**	**Numbers of municipalities**
Municipality 1	0 to 2,000	194
Municipality 2	2,001 to 20,000	820
Municipality 3	20,001 to 200,000	93
Municipality 4	More than 200,001	4

Table 2. Classification of municipalities according to census population data

Infoterra LaND25 dataset contains geometric and attribute information on 26 different types of LU. Six LU classes are distinguished as residential areas (both in urban and rural areas) which accounts for 6% of the total area in BW, compared to 88% as vegetation purposes. It is assumed that people live only in the residential areas. Therefore, population data will be disaggregated in the residential classes. Other categories of LU, such as industry or public buildings may contain population, but water and artificial areas do not contain any population.

The investigation with LaND25 (Infoterra 2007) and CLC2000 data (UBA, DLR et al. 2004) show that industrial areas are not correctly defined within the LaND25 data, many of such types are demarcated as residential LU data. On the other hand, CLC2000 provides more precise information regarding the 42 LU classes. Therefore, using GIS overlay operations, these industrial zones are erased from the LaND25 data. As the Green Heat project also investigates the potential of local renewable sources, this study focuses on the residential areas built before 1990. Therefore, buildings built after 1990 consequently show little potential and are excluded from the scope of the study. As CLC2000 also describes the LU changes between 1990 and 2000, these areas are erased from the LaND25 data. In the end, about 2,366 km² of residential LU areas are represented in about 57,000 LU units.

It is observed that not all the municipalities are occupied with every type of residential LU. In general, municipalities having higher population number are occupied with maximum LU types. But to perform regression analysis of the different types of LU, it is important that the municipalities should have the representative LU types. Therefore, the six residential LU are reclassified into four types (Table 3).

Infoterra LaND25 (LU)	Description	Proposed LU
LU1	Extremely dense urban	Code 12
LU2	High buildings in extremely dense urban	
LU3	Dense urban	Code 34
LU4	High buildings in dense urban	
LU5	Urban fabric	Code 5
LU6	Village and suburban	Code 6

Table 3. Reclassification of Infoterra LaND25 residential LU classes

In this step, municipal population data (StaLaBW, 2007) is overlayed on the LaND25 residential LU classes to prepare an inventory dataset of municipality-wide LU information and population. This dataset is the basis for further regression analysis and modelling.

b) Development of regression model

Linear regression analysis is used for disaggregation of population data. The main objective of the linear regression is to describe the linear link existing between the dependant variable and of one or more independent variables. The dependant variable in the regression equation is modelled as a function of the independent variables, corresponding parameters, named "constants", and an error term. This error term represents an unexplained variation in the dependant variable.

It is described earlier that the population distribution within the municipalities are not homogeneous, but is expected to relate it to LU types since LU properties can be one factor in the distribution process (Flowerdew and Green, 1989). Therefore, considering the statistical data on the number of population of each municipality (dependent variables) and the area of different types of residential LU (independent variables), linear regression models are built for every category of municipality to estimate the coefficients for each LU type. Finally, the coefficients can be allocated to the LU classes to determine the number of population. Following regression equation is used for each category of municipality:

$$P_i = \sum_{j=1}^{n} (b_j \times x_j + \varepsilon i)$$

Where, P_i: predicted population of the municipality i, x_j: residential area of the LU type j, εi: errors, b_j: coefficient determined for residential LU type j.

For the municipality category 1, the regression model is run for the three LU classes in 194 municipalities (Table 2 and Table 3). The multiple coefficient of determination R square = 0.69 indicates that the “goodness of fit” of regression line is middle high (Table 4). Whereas the regression model for the municipality category 3 is run for four LU classes in 93 municipalities and the R square = 0.87 indicates that the “goodness of fit” for regression model is very high. But because of the very limited number of represented municipalities in category 4, no regression model is applied. So, disaggregation of data in these municipalities is not performed. Table 4 and Table 5 summarise the results of the regression models.

Proposed LU	**Infoterra LU**	**Coefficient (weight/m²)**		
		Municipality 1	Municipality 2	Municipality 3
Code 12	LU1	-	-	0.078166
	LU2	-	-	0.078166
Code 34	LU3	0.004616	0.013765	0.013696
	LU4	0.004616	0.013765	0.013696
Code 5	LU5	0.002645	0.006549	0.009145
Code 6	LU6	0.001991	0.002849	0.003376

Table 4. Estimated population densities coefficient for each residential LU classes in different municipality categories

Linear Regression statistics			
	Municipality 1	Municipality 2	Municipality 3
Intercept/error	267.3030702	405.8836834	791.357419
Multiple R	0.831805	0.891197	0.934276
R Square	0.691899	0.794232	0.872872
Adjusted R Square	0.687034	0.793476	0.867094
Standard Error	329.6397	1821.173	9213.72
Observations	194	820	93

Table 5. Linear regressions statistics for different municipality categories

Once the weight of every LU class as well as the error for each category of municipality has been calculated by regression models, the population number in each LU class can thus be estimated using the above regression equation.

c) Use of scaling techniques and estimation of population

The estimated population can be a further adjusted using scaling technique. There are two reasons to modify the estimation of the previous steps. First, the distribution of estimated errors from the regression is large. Secondly, it is obvious that population distribution is not explained solely by LU type e.g. the number of residential population living in far urban centre may be less than population in residential areas near to urban centres and that transportation system, hydrology, natural environment and socio-economics settings all may influence population distribution. Therefore, in order to reduce the error distribution and the influences of above parameters, scaling techniques is applied (Flowerdew and Green, 1989), (Yuan et al., 1997).

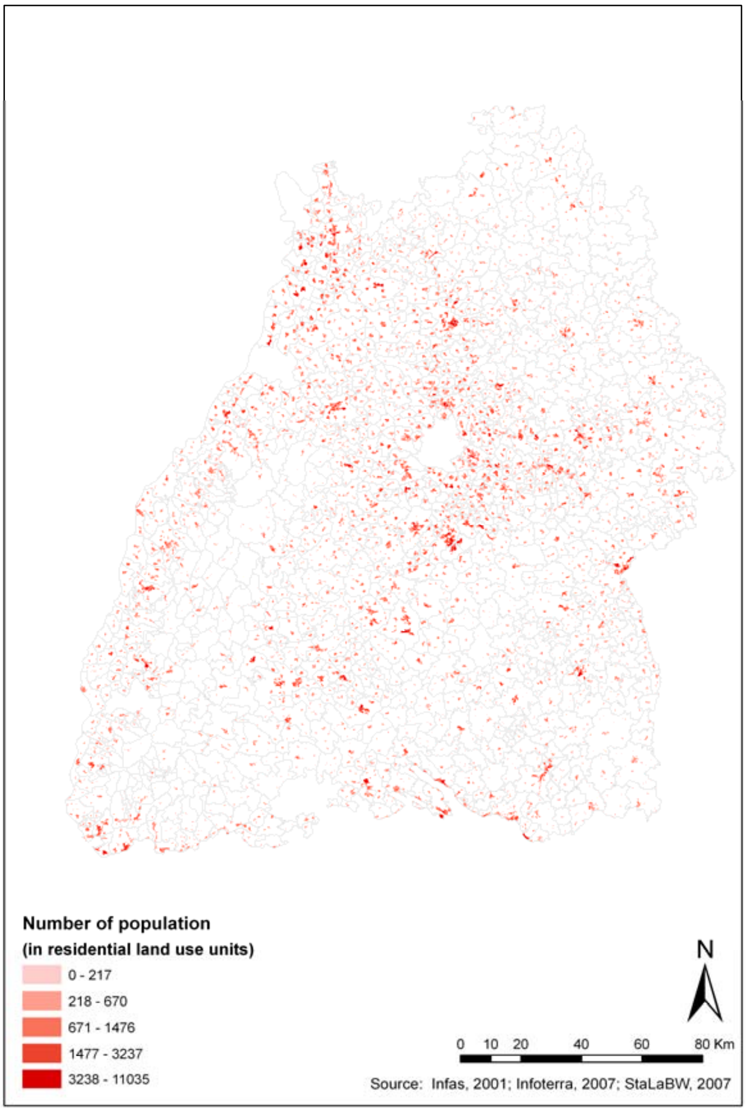

Figure 3. Disaggregation of statistical population data into the residential LU units in Baden-Württemberg, Germany

Scaling technique is straightforward. It is assumed that the statistical population Y_i at the municipality are highly reliable and that the estimated population of the municipality can be scaled to more refined estimation. Therefore, following mathematical equation is used to correct the population in each type of the LU in different municipalities:

$$b_{ij} = \frac{Y_i}{P_i} \times b_j \times x_j$$

Where, b_{ij}: corrected population for residential LU *j* within the municipality *i*, Y_i: statistical population of the municipality *i*, P_i: predicted population of the municipality *i*, x_j: residential area of the LU type *j*, b_j: coefficient determined for residential LU type *j*.

The homogeneously distributed statistical population data at the municipality level is, therefore, disaggregated into a finer scale of analysis, at LU units. The number of population within the different LU varies significantly, depending on the size of LU, type of LU and municipality (Figure 3). Number of population in the different LU units varies significantly from 0 to about 11,000 inhabitants. In general, population number is higher in urban residential LU than that of village and suburban LU. This disaggregated population provides valuable input for the assessment of hot water demand distribution in BW.

4. Calculation of hot water demand

The disaggregated population is considered as the basis for calculation of hot water demand by considering the specific assumptions on the consumption of hot water. Other parameters such as water heating systems or occupant behaviour characteristic are not considered because the quantification of such factors is very difficult. Therefore, a general assumption to estimate the hot water demand is based on the number of inhabitants. Most of the demand is due to personal hygiene, mainly showering and bathing, and this demand is rather proportional to the population. (Stadtwerke Hildesheim, 2008) indicates that the average hot water consumption in Germany amounts 80m³ to 100m³ per inhabitant and per year i.e. 750 kWh/(inh.yr) to 1,070 kWh/(inh.yr). So the average value of 1,000 kWh/(inh.yr) is chosen to calculate the hot water demand (Stadtwerke Klagenfurt STW and Energie Klagenfurt EKG, 2008), (Energiespar-Rechner, 2008).

The hot water demand in each LU type of every municipality can thus be calculated using the following equation:

$$\mathrm{HWD_{ij}}\left[\frac{\mathrm{kWh}}{\mathrm{a}}\right] = \mathrm{w}\left[\frac{\mathrm{kWh}}{\mathrm{inh.a}}\right] \times \mathrm{y_{ij}}[\mathrm{inh}]$$

Where, $\mathrm{y_{ij}}$: number of population in the residential LU *j* in municipality *i*, w: hot water consumption coefficient.

Finally, hot water demand in each of the LU unit within the municipalities is calculated. The hot water demand is linearly correlated to the population. Figure 4 shows that the hot water demand is higher in highly populated LU units.

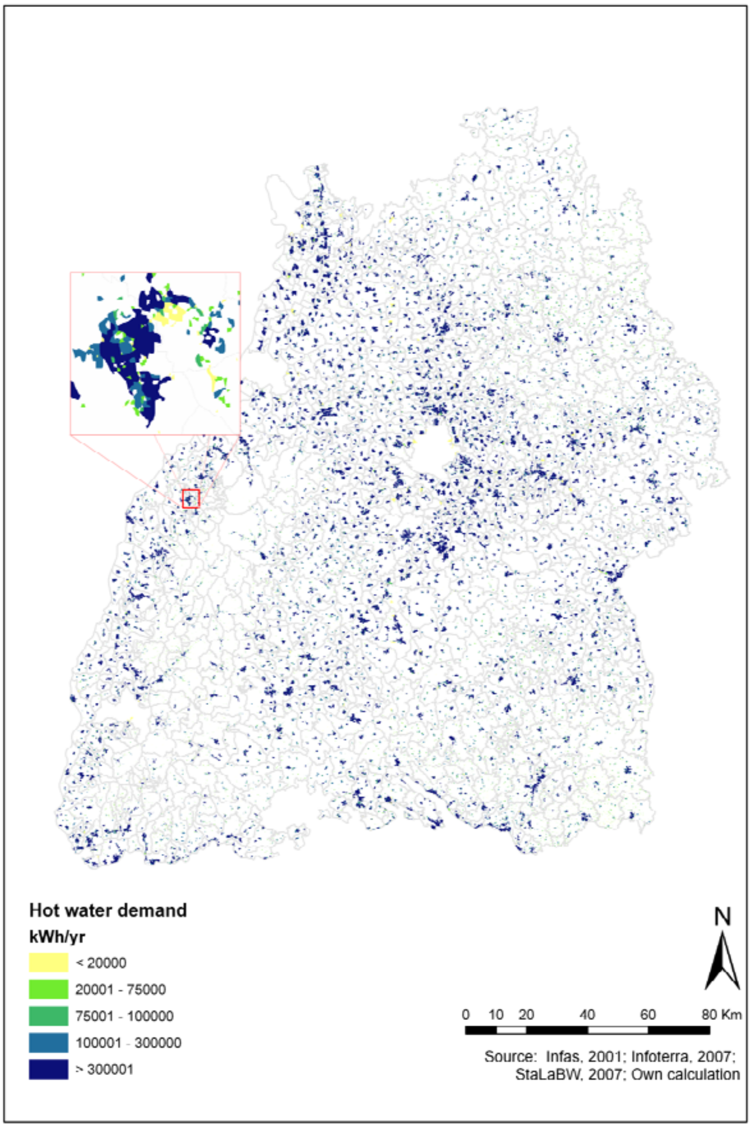

Figure 4. Hot water demand in the residential LU units in Baden-Württemberg, Germany

5. Discussion and Conclusion

The disaggregation methodology realized in the study follows a systematic approach. For instance, the regression analysis is performed for three different categories of municipalities; therefore, the regression models are optimised and provide with better regression results. The residential hot water demand gives a very detailed spatial dimension of heat demand for the entire state of BW. By combining the space heating demand carried out in BW (Murshed et al., 2009), this study would provide clear overview of the heating situation in BW. It would help the policy makers and energy companies to plan the cost effective and efficient heating solutions. Such analysis being applied in BW can also be relevant to other regions in Germany.

In disaggregation of population, four largest cities, namely Stuttgart, Karlsruhe, Mannheim and Freiburg im Breisgau are not considered. The reason is the lack of such representative cities where statistical regression analysis can be meaningfully performed. With an accuracy usually comprised between ±25%, the disaggregation of the population showed very good results in compared to the results of other similar studies (Wu et al., 2005). Nevertheless, as the hot water demand is calculated as proportional to the population, the reliability of the results regarding the hot water demand varies accordingly. In the end, this demand represents less than 15% of the total residential hot water demand in the municipalities. So the disparity of the results remains quite limited.

The development of hot water demand methodology and its application in BW revealed an open field of research and great investigation opportunities regarding data disaggregation. It is applied only to disaggregate population data, but other kind of socio-demographic and energy-related data can be disaggregated as well. Moreover, if higher resolution of satellite data with better interpretations of LU or the cadastre data (e.g. ALK data) are available, the disaggregation methodology can be further improved.

Acknowledgement

The author is grateful to EnBW for providing financial support to carry out the Green Heat Project as well as permission to publish this paper. Project members, especially Sébastien Girard, Anthony Methenier and Jacques Ghisgant are also acknowledged.

References

BMVBS (2007) Wohnen und Bauen in Zahlen 2007. Bonn, Bundesministerium für Verkehr Bau und Stadtentwicklung.

DESTATIS (2007) Statistisches Jahrbuch 2007 für die Bundesrepublik Deutschland. Wiesbaden, Statistischen Bundesamt Deutschland.

EICHER, C. L. & BREWER, C. A. (2001) Dasymetric Method and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science,* 28**,** 125-138.

ENERGIESPAR-RECHNER (2008) Online Berechnung Amortisation von thermischen Solaranlagen, Lohnt sich die Anschaffung von thermischen Solaranlagen für Trinkwassererwärmung?

FLOWERDEW, R. & GREEN, M. (Eds.) (1989) *Statistical methods for inference between imcomparable zonal systems,* Newyork, Taylor and Francis.

HUANG, Z. & OTTENS, H. (2007) A Doubly Weighted Approach to Urban Data Disaggregation in GIS: A Case Study of Wuhan, China. *Transactions in GIS,* 11(2)**,** 197 - 211.

INFAS (2001) INFAS GEOdaten: Das DataWherehouse.
INFOTERRA (2007) Infoterra LaND25 - Daten Baden-Württemberg & Puffer. Infoterra GmbH.
LANGFORD, M. (2006) Obtaining population estimates in non-census reporting ones: An evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems,* 30**,** 161-180.
LANGFORD, M., MAGUIRE, D. J. & UNWIN, D. J. (Eds.) (1991) *The Areal Interpolation problem: estimating population using remote sensing in a GIS framework.* London, Longman.
MENNIS, J. (2003) Generating Surface Models of Population Using Dasymetric Mapping. *The professional Geographer,* 55**,** 31-42.
MURSHED, S. M., GIRARD, S. & GHISGANT, J. (2009) Development of a top-down approach to calculate residential space heating demand in Baden-Württemberg, Germany. *3rd ICA Workshop on Geospatial Analysis and Modeling.* Gävle, Sweden.
STADTWERKE HILDESHEIM (2008) Wasser Eenergiespartipps. Hildesheim.
STADTWERKE KLAGENFURT STW & ENERGIE KLAGENFURT EKG (2008) Nachtspeicherheizung/Warmwasser. Klagenfurt.
STALABW (2007) Fläche, Bevölkerung - Daten zu Baden-Württemberg. Stuttgart, Statistisches Landesamt Baden-Württemberg, Germany.
UBA, DLR & DFD (2004) CORINE Land Cover 2000 - Daten zur Bodenbedeckung Deutschland.
WU, S.-S., QIU, X. & WANG, L. (2005) Population Estimation Methods in GIS and Remote Sensing : a review. *GIScience and Remote sensing,* 42**,** 58-74.
YUAN, Y., SMITH, R. M. & LIMP, W. F. (1997) Remodeling census population with spatial information from Landsat TM imagery. *Computer, Environment and Urban Systems,* 21**,** 245-258.