Population Distribution Modelling for Calibration of Multi-Agent Traffic Simulation

Christian Kaiser[1], Mikhail Kanevski[2]
[1] Institute of geography, University of Lausanne, Switzerland
[2] Institute of geomatics and analysis of risk, University of Lausanne, Switzerland

ABSTRACT

Estimating the population distribution is an important issue in multi-agent traffic simulation, and also in many other fields related to geography. This paper presents a methodology to find an exact location for each agent using aggregated population data, the location of residential buildings, and the buildings volume for estimating the distribution itself. The presented approach is validated for a case study using high-resolution census population data. The comparison between the estimated population distribution and the real distribution allows the assessment of the quality of the approach. The contribution of each information layer is estimated. The resulting population distribution matches quite well the real distribution, but shows also important zones with over- or under-estimation of population.

1. INTRODUCTION

Traffic planning is an important issue in modern cities and agglomerations. Traffic is essential for economy, but has also negative effects like noise, air pollution or injuries caused by accidents. Inefficient traffic, like the occurrence of traffic jams, increases these negative effects and results in time loss for involved people or reduces the overall efficiency of economy. Individuals using the transportation system are not concerned about the system itself, but only about their own benefit. As transportation capacity on a network is limited, competition for the use of this offer arises in densely populated places (Balmer, 2007). Transportation planning is an important tool for offering an optimal network for the needs of the economy and the people by limiting the negative effects of traffic.

Traffic is generated by people for accomplishing their activities. The traffic planning should ideally integrate the reasons for the mobility behaviour at an individual level. This would enable the planner to monitor each person and get information about the traffic volumes at a given moment in time, the modal split and the reasons for the modal choice or the activity chain (Balmer, 2007). Such detailed real world data do of course not exist because of privacy issues. In traffic planning, tools and techniques for dealing with incomplete data are therefore essential.

Micro-simulation is an important tool in traffic simulation (TS) (Balmer, 2007; Vovsha, Petersen, & Donnelly, 2002; Bowman, Bradley, Shiftan, Lawton, & Ben-Akiva, 1999; Bhat, Guo, Srinivasan, & Sivakumar, 2004). TS can provide us with a precise spatiotemporal image of the real traffic. TS can also be used to understand interactions between the urban structure and the real traffic. It allows estimating more accurately the transportation duration with respect to different traffic situations during a day. And it can be used to simulate the car dependence for different populations by comparing the transportation duration for private and public transportation means. It is also possible to integrate the individual decision-making process in the TS.

For a multi-agent traffic simulation, it is therefore necessary to know with the best possible precision the location of each person for the population under study. Generally, we do not know this location and we have to estimate it based on aggregated data coming for example from the population

census. This problem can be considered as a variant of the modifiable area unit problem (MAUP) (Openshaw, 1984), as this population distribution problem can be considered as the estimation of the population density for an arbitrary small region. Dasymetric mapping is one possible method where the distribution of the aggregated data within the unit of analysis is estimated by using additional information on how the data are potentially distributed (Wright, 1936). In this paper, we use the dasymetric mapping for locating each individual in the population for a multi-agent traffic simulation in the agglomeration of Lausanne, in Switzerland. We evaluate the quality of this mapping and determine the contribution of LiDAR (Light Detection And Ranging) elevation data.

2. PREVIOUS WORK

The methods for estimating the population distribution or the population density for a given area are numerous. Wu, Qiu and Wang (2005) separate all the different approaches into two categories: the methods of area-to-point interpolation and the statistical models. Area-to-point interpolation allows solving the zone transformation problem where we have to estimate a variable known only for a set A of spatial units for another generally bigger-scaled set B of spatial units. Such a transformation involves the estimation of the spatial distribution of the measured phenomenon. For each geographic unit of set A, we need the exact value for the variable, which can come for example from a census. Statistical models try to apply urban geography theories in order to estimate the population distribution. These approaches try to establish a (statistical) link between the population and another variable (built area, land use, satellite imagery, etc.).

Area-to-point interpolation allows the estimation of the population inside a spatial unit smaller than the one used by the census. We call source zone the set of units where we know the population, and target zone the set of usually smaller units for which we want to estimate the population (Lam, 1983). Martin (1989) proposes a kernel-based interpolation method using the source zone centroid as a control point. He uses a distance weighting function for estimating the population for each node of a regular grid. This method supposes that each geographic unit is more or less symmetric, which is in reality rarely the case. This method, as other methods using the centroid as a control point, cannot guarantee that he total population is the same after interpolation, which is a big inconvenience (Wu, Qiu and Wang, 2005).

Tobler's pycnophylactic interpolation is a well-known area-to-point interpolation method (Tobler, 1979). A smooth, continuous density function is computed in space taking into account the neighbourhood and guarantees the respect of the total population after interpolation. The estimated surface approximates the neighbourhood mean. Kyriakidis (2004) presents a geostatistical approach for the area-to-point interpolation. His method is a generalisation of Tobler's pycnophylactic interpolation (Tobler, 1979). This geostastistical framework allows the estimation of the value at each point in space using the area values and with respect to the total population. Kyriakidis (2004) considers the area-to-point interpolation as a special case of change of support and refers to Gotway and Young (2002). The proposed method is a special case of Kriging (Matheron, 1971), even if Kriging has more often been applied to classic point-to-point interpolation problems.

Wright (1936) has developed the «dasymetric method» which uses auxiliary information for restricting the domain where the uniform distribution can be applied. Poulsen and Kennedy (2004) define this approach as follows: *«Dasymetric mapping involves estimating the distribution of aggregated data within the unit of analysis, by adding additional information that provides insights on how these data are potentially distributed.»* The idea is to ignore regions without population, which is the same as to make a binary distinction between presence and absence of population. Using Geographical Information Systems (GIS), this approach has become very widespread and easy to implement. The method has also been applied to the non-binary case, for example by Maantay,

Maroko and Herrmann (2007) who use a cadastral-based expert system, or by Langford, Maguire and Unwin (1991) who use a regression analysis in order to refine the density in populated places. More case studies, some with variants of the dasymetric approach, are known in the literature, see e.g. Wu, Qiu and Wang (2005) for some of them. The advantage of the dasymetric interpolation is the ease of integration of several factors in order to get a probability map characterising the population distribution.

However, it is often unclear how accurate the result of these methods is. The pycnophylactic interpolation or the geostatistical approach give a smooth representation over the whole region. Population is typically distributed in town centres and around, and there are generally clear limits between inhabited zones and agricultural land or forest. For getting a more accurate calibration for a traffic simulation, we can introduce some auxiliary information. For example, we can include a validity domain consisting of the residential buildings. We can also include other information on the local population density. It is often unclear if such auxiliary information improves the population distribution estimation significantly, and if yes, by which amount.

3. METHODOLOGY

Our approach for estimating the population distribution can be seen as a variant of dasymetric interpolation. Two different kinds of auxiliary information are used: a validity domain v and a density layer ρ.

The validity domain defines for each location i in our study region if at this point somebody could potentially live:

$$\begin{aligned} v_i &= 1 \quad \text{if } i \text{ is lying inside a populated zone} \\ v_i &= 0 \quad \text{otherwise} \end{aligned} \tag{1}$$

The validity domain can either be a polygon layer or a high-resolution raster layer. In our case, the validity domain is defined as a polygon layer consisting of the residential buildings extracted from a topographic map; the type of building has been determined with the help of a land use map.

The density layer ρ represents an estimate of the normalised population density at locations j for the whole region:

$$\rho_j \propto p_j \qquad \text{with } \rho_{min}=0 \text{ and } \rho_{max}=1 \tag{2}$$

The density layer is typically a raster layer, or a combination of several raster layers, that we consider as being proportional to the real population density. In our case, the building volume obtained from the LiDAR elevation data and the topographic map provides this estimate. The LiDAR technology allows estimating the building heights for a whole, even quite large region at high resolution. In combination with the building area from the topographic map, we can deduce the building volume. In this example, the population is distributed proportionally to the volume of the residential buildings.

The validity domain and the density layer can then be used for estimating the exact location of the whole population (for each agent). Table 1 describes each step of the algorithm. The algorithm determines the exact location of each agent based on the validity domain and the density layer. However, due to some random components, the result will be slightly different for each run. The selection of the exact location is based on a random process. If this exact location i is lying inside the validity domain, the location i is only retained if the density layer value ρ_i at the same location is

bigger or equal to a random number *pval*. Remind that the density layer ρ is normalised to have only values between 0 and 1, so the random number *pval* is also lying between 0 and 1. The algorithm yields for each run a slightly different result; this is a reasonable approach as the exact population distribution is not known and each of the result is a possible solution to the given problem. For instance, if 2 cells have both a density layer value ρ_i=0.1, and 3 people should be distributed in these 2 cells, there is no clear reason that one of the cells should contain 2 people, and the other only 1 person. A random component is therefore an appropriate approach to handle this uncertainty.

Require:	A bounding box *B* enclosing our study region, a polygon layer with our validity domain *v*, a raster layer with the density estimation ρ
1:	Initialize an empty population array `pop`
2:	**repeat**
3:	get a random location `i` inside `B`
4:	**if** `i` inside `v` **then**
5:	get a random probability value `pval` such as $0 < pval \leq 1$
6:	get ρ at location `i`
7:	**if** `pval` $\leq \rho$ **then**
8:	add location `i` to `pop`
9:	**end if**
10:	**end if**
11:	**until** the whole population is located

Table 1. *The population distribution algorithm.*

In order to evaluate the accuracy of the population distribution, and the contribution of the building heights data, we estimate the distribution based on population data at the level of the commune. Once the individual locations estimated, the data is aggregated according to a regular grid of 100 metres resolution for which the same population data is available. We are then able to compare the obtained results with the real population distribution known from census data. Several runs of the algorithm allow further to estimate the stability of the result. This is a necessary step as the algorithm contains random components.

4. CASE STUDY: THE LAUSANNE AGGLOMERATION

70 communes compose the Lausanne agglomeration, in Switzerland. About 300'000 people are living in the agglomeration. Population is distributed in a very unequal manner, with the city of Lausanne and its neighbouring communes having a much higher population density. Generally, population is concentrated in the commune's centre.

We have estimated the population distribution for the whole agglomeration at an individual level using three methods: (1) a uniform distribution inside each commune, without validity domain or density layer; (2) a uniform distribution with the residential buildings as validity domain, but without density layer; (3) a uniform distribution with the residential buildings as a validity domain, and the building volume as a density layer. For each method, the population has been aggregated to the hectometric grid allowing a comparison with the gridded census data. Figure 1 shows a small portion

of the study area in the centre of the agglomeration, with the census data at left, and the estimated population data based on the validity domain and the density layer at right.

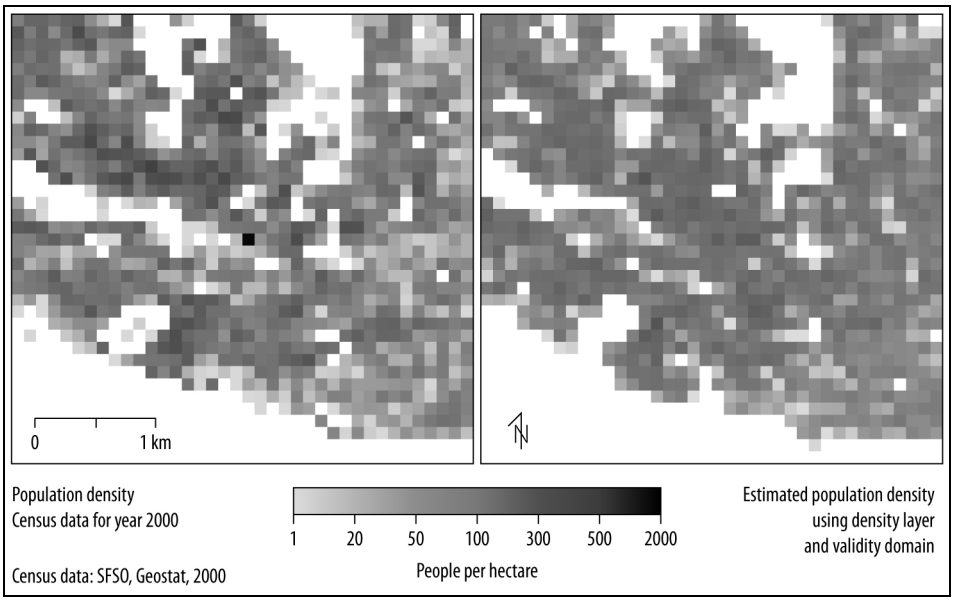

Figure 1: Comparison between real population density (left) and estimated density using the residential buildings as validity domain and the buildings volumes as density layer (right) for a small area in the centre of the agglomeration of Lausanne.

Pearson's correlation coefficient allows the assessment of the accuracy of the different methods, and an estimation of the contribution of the auxiliary information. Table 1 shows the different correlation coefficients. Comparing with the dasymetric interpolation using only a validity domain, which is identical to Wright's (1936) binary approach (method 2), the use of a density layer (method 3) improves considerably the correlation coefficient. The simple uniform distribution (method 1) shows a clearly unsatisfactory result.

	Pearson's correlation coefficient
Method 1: Uniform distribution	0.374
Method 2: Uniform distribution with validity domain	0.561
Method 3: Uniform distribution with validity domain and density layer (presented approach)	0.603

Table 2: Correlation coefficients between the real population values and each of the simulated population values.

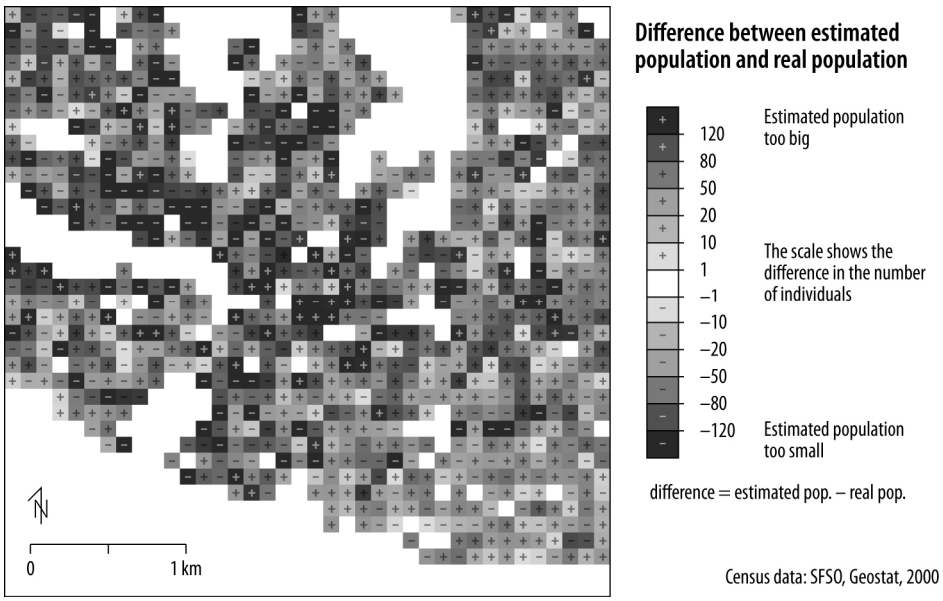

Figure 2: Difference between the simulated population distribution (as in figure 1 right) and the real population distribution (as in figure 1 left) for the same small area in the centre of Lausanne as in figure 1.

Even with a quite big number of included data in our estimation process, the resulting distribution does not correspond entirely to the reality. Figure 2 shows the differences between the estimated and real population values for a small area in the centre of Lausanne. The central commune is the biggest in the whole agglomeration, and the uncertainty for estimating the population distribution is therefore the highest. There is no clear visible pattern in figure 2. Analysis of the whole agglomeration shows that we have a slight over-estimation of the population for large zones. Figure 3 shows the frequency histogram for the differences between the estimated population and the real population. It confirms the over-estimation for large zones. These zones are often residential districts with a relatively low density. These districts with probably a more fortune population, the building volume occupied by person is higher than the overall mean, which results in an over-estimation of the population in these zones. We have also an over-estimation in Lausanne's city centre, which is due to the presence of a high number of offices, commercial activities and other, e.g. administrative buildings being partly residential. The data used for estimating the population does not contain any information if a building is only partially residential or not. Under-estimated zones are present mainly in the western periphery of the city of Lausanne. These are residential districts with slightly poorer population classes where the population density per volume is higher than expected. Some cells containing very high population values (up to 1842 people per hectare) cannot be estimated correctly. Some extremely under-estimated cells (not shown in figure 3) counterbalance the numerous cells with slightly over-estimated population.

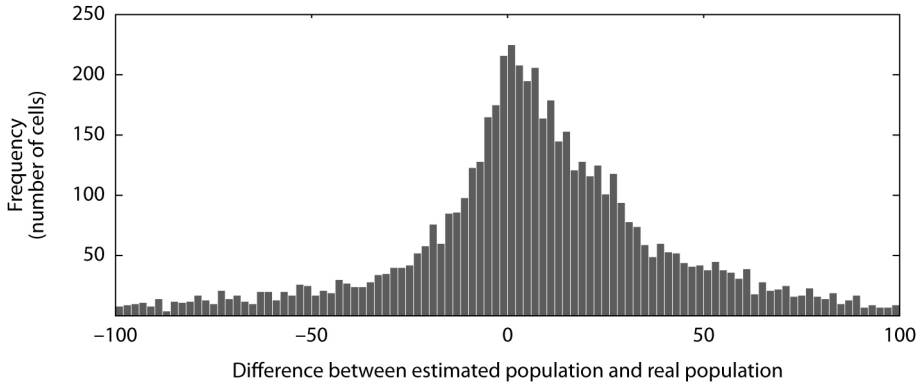

Figure 3: Frequency histogram of the population estimation errors for the whole agglomeration.

As already explained, the presented population density estimation method has a random component; the results vary therefore from one run to another. Repeated simulations allow estimating the variations between different runs. We have run 100 estimations and computed the variation between the different runs for each cell. The chosen measure for the variation is the standard deviation. Figure 4 shows the frequency histogram of the standard deviation for the population values of the 100 estimations computed for each cell. The mean value for all standard deviations inside the validity domain is 4.4, the median is 3.6 people per hectare; the mean population value for the census data is 37.9 people per hectare. For 93.5% of all cells, the standard deviation is less than 10. The variations between different runs can be considered as quite low. If we compute the variations in Pearson's correlation coefficient as presented in table 1, between the 100 estimated population distributions and the real population distribution, the differences are only minimal.

The analysis of the variations between different runs of the population estimation allows the assessment of the stability of the procedure. If the variations are high, the population distribution is unstable and one should consider another methodology for estimating the population distribution, or get better initial data. This analysis does of course not give any estimate on how well the population is distributed, but it can be done without knowledge of the real distribution. It should be performed for each population distribution estimation as a proxy for the estimation stability.

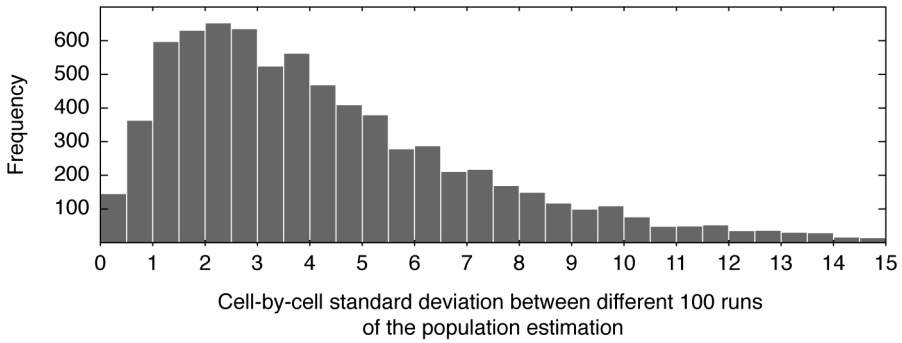

Figure 4: Histogram for the standard deviation for each hectare between different runs of the population distribution estimation.

5. DISCUSSION

Our analysis has clearly shown that it is necessary to use auxiliary information for a more or less realistic population distribution. This is confirmed by Langford et al (2008) which have studied the impact of the urban population distribution modelling on service accessibility: *«...it has been shown that the choice of population distribution model [...] can exert a significant influence on outcomes.»* Even with the most sophisticated model, it is important to be careful and perform a sensibility analysis for the model used. However, it has been shown that the building heights acquired by LiDAR elevation data can considerably improve the modelling of the population distribution.

The maps produced for validation (figure 2) can be used for localising zones with under- or over-representation of population in comparison with the auxiliary information included in the population estimation process.

The use of population density models issued from urban geography may improve the result. A lot of researchers have noted a decrease of the population density from the city centre toward the suburbs. Clark (1951) has described this relation mathematically (Wu, Qiu and Wang, 2005). However, this simple concentric model does not consider functional differences between different city districts.

Additional research is necessary for establishing a link between the urban morphology and the population density. The use of socio-economic data, e.g. poverty, may also improve the result's quality.

Even if the presented approach shows promising results for the population distribution estimation, more work should be done for assessing the quality and the errors of the estimation. Other indices than Pearson's correlation coefficient computed on a cell-by-cell basis could be considered. One candidate is the Fuzzy kappa index used for estimating the accuracy of cellular automata simulations (Hagen, 2003, Hagen-Zanker, 2006). This index has been created for categorical data and considers differences in category and location. An adaptation to non-categorical data is straightforward.

A correlation coefficient based on a cell-by-cell basis of around 0.6 is a quite good result for the population distribution estimation. However, the fact that for large zones, the population is over-estimated, and some highly populated cells are missed could lead to some issues in a multi-agent traffic simulation. Some artefacts might be the result in the zones where population has been over-estimated, for example traffic jams at unlikely places, or agendas not corresponding to reality. However, no work has yet been done in order to estimate such effects. Another important issue not assessed until now is the spatial resolution of the initial population data used for downscaling. If we don't have population data per commune, but only for some bigger regions, the accuracy of the population estimation will decrease.

ACKNOWLEDGEMENTS

This work has been supported by the Swiss National Foundation, through the projects «Urbanization Regime and Environmental Impact: Analysis and Modelling of Urban Patterns, Clustering and Metamorphoses» (nº 100012-113506) and «GeoKernels»: Kernel-Based Methods for Geo- and Environmental Sciences, phase 2 (nº 200020-121835).

REFERENCES

Balmer, M., 2007 Travel demand modeling for multi-agent transport simulations: algorithms and systems. PhD thesis. Swiss Federal Institute of Technology, Zurich, Switzerland. Doi:10.3929/ethz-a-005429518.

Bhat, C.R., Guo, J.Y., Srinivasan, S. and Sivakumar, A., 2004 A comprehensive econometric microsimulator for daily activity-travel patterns. Transportation Research Record 1894, 57–66.

Bowman, J.L., Bradley, M., Shiftan, Y., Lawton, T.K. and Ben-Akiva, M.E., 1999 Demonstration of an activity-based model for Portland. World Transport Research 3, 171–184.

Clark, C., 1951 Urban population densities. Journal of the Royal Statistical Society 114, 490–496.

Gotway, C.A. and Young, L.J., 2002 Combining incompatible spatial data. Journal of the American Statistical Association 97(458), 632–648.

Hagen, A., 2003 Fuzzy set approach to assessing similarity of categorical maps. International Journal of Geographical Information Science, 17(3), 235–249.

Hagen-Zanker, A., 2006 Map comparison methods that simultaneously address overlap and structure. Journal of Geographical Systems, 8(2), 165–185.

Kyriakidis, P.C., 2004 A geostatistical framework for area-to-point spatial interpolation. Geographical Analysis 36(3), 259–289.

Lam, N., 1983 Spatial interpolation methods: A review. The American Cartographer 10(2), 129–149.

Langford, M., Maguire, D., and Unwin, D., 1991 The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In: Masser, I. and Blakemore, M. (Eds), Hand ling Geographic Information: Methodology and Potential Applications. Longman, London, UK.

Martin, D., 1989 Mapping population data from zone centroid locations. Transactions of the Institute of British Geographers 14(1), 90–97.

Maantay, J. A., Maroko, A. R., and Herrmann, C., 2007 Mapping population distribution in the urban environment: The cadastral-based expert dasymetric system (CEDS). Cartography and Geographic Information Science 34(2), 77–102.

Matheron, G., 1971 The Theory of Regionalized Variables and its Applications. Les cahiers du centre de morphologie mathématique de Fontainebleau. Ecole Nationale Supérieure des Mines, Paris.

Openshaw, S., 1984 Ecological fallacies and the analysis of areal census data. Environment and Planning A 16(1), 17–31.

Poulsen, E., and Kennedy, L., 2004 Using dasymetric mapping for spatially aggregated crime data. Journal of Quantitative Criminology 20(3), 243–262.

Swiss Federal Statistical Office SFSO, 2005 Recensement fédéral de la population 2000 (recensement de la population et des ménages, recensement des bâtiments et des logements). Technical report, SFSO Geostat, Neuchâtel.

Tobler, W., 1979 Smooth pycnophylactic interpolation for geographical regions. Journal of the American Statistical Association 74(367), 519–530.

Vovsha, P., Petersen, E. and Donnelly, R., 2002 Microsimulation in travel demand modeling: lessons learned from the New York best practice model. Tranportation Research Record 1805, 68–77.

Wright, J. K., 1936 A method of mapping densities of population: With Cape Cod as an example. Geographical Review 26(1), 103–110.

Wu, S.-S., Qiu, X., and Wang, L., 2005 Population estimation methods in GIS and remote sensing : A review. GIScience and Remote Sensing 42, 80–96.