Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200157

Generation of Fine Grained Demographic Information for Epidemiological Analysis

Timo WOLTERS [a,1], Oke WÜBBENHORST [a], Christian LÜPKES [a] and Andreas HEIN [b]

[a] *Division Health, OFFIS - Institute for Information Technology, Escherweg 2, Oldenburg, Germany*

[b] *Department of Health Services Research, Carl von Ossietzky University Oldenburg, Oldenburg, Germany*

Abstract. Cancer risks may be influenced by local exposures such as working conditions or nuclear waste repositories. To find influences, local accumulations of cancer rates are used, for which finely granulated data should be utilized. In particular, high-resolution demographic data for a reference population are important, but often not available for data protection reasons. Therefore, estimation methods are necessary to approximate small-scale demographic data as accurately as possible. This paper presents an approach to project existing epidemiological and public data to a common granularity with respect to attribute characteristics such as place of residence, age or smoking status to allow for analyses such as local accumulations and consistently falls below an average relative error of 5%.

Keywords. Health Data Science, Data Analysis, Population Synthesis, Microsimulation

1. Introduction

The risk of cancer can be increased by the degradation of nuclear waste repositories and the associated radiation exposure of the immediate environment. Local accumulations are a type of analysis in cancer epidemiology to correlate geospatial impact with cancer risk and requires finely granular data from cancer incidences and a reference population. In epidemiological cancer registries, the reference population is often more coarsely granular than the incidences due to data protection, making an approximation of the population at the incidence level desirable. This paper presents an approach for the generation of fine-grained data using iterative proportional fitting (IPF) combined with cellular automata (CA) for exemplary application in analyses of epidemiological cancer registration (ECR) in Lower Saxony and evaluates the accuracy and reliability of the generation.

In the ECR, data is evaluated pseudonymised by the registry unit, which as a result of the analysis for the verification of a suspicion can request the underlying not pseudonymised data from the confidentiality unit [1]. In order to form a suspicion, exact results are therefore not necessary at first, since the analysis is finally carried out on the

[1] Corresponding Author: Timo Wolters, OFFIS - Institute for Information Technology, Escherweg 2, Oldenburg, Germany; E-mail: timo.wolters@offis.de.

data of the confidentiality unit. However, too large an approximation error leads to false suspicions and should therefore be minimized.

Incidences are present in the ECR for Lower Saxony in a geometric grid with an edge length of 1km (1km grid) while the reference population is present at municipality level. Figure 1 shows the difference between the municipality level and a 1km grid, while the actual distribution of the population is shown in a 100m grid using the municipality of Oldenburg. The 1km grid has a much finer resolution than the municipality and better reflects the spatial distribution of the population, especially with respect to the temporal stability of a grid in contrast to frequently changing administrative boundaries of municipalities [1].

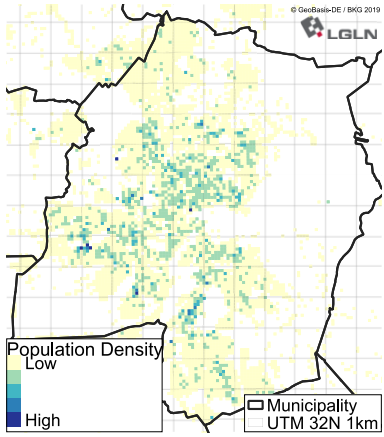


Figure 1. Granularity comparisons between a municipality, an 1km grid and an 100m grid for Oldenburg in Lower Saxony.

After Section 2 has outlined related work on population synthesis and microsimulations, Section 3 describes our approach to generate fine-granular data. Section 4 explains the implementation details and Section 5 evaluates the generation of the population, while Section 6 describes future work and summarises this contribution.

2. Related Work

Ballas et.al. investigate in [2] location-based microsimulation systems for population synthesis. Population synthesis usually combines fine granular census data with coarse granular data to approximate a synthetic population between two censuses or to add missing attributes. Population synthesis can be divided into population reconstruction, probabilistic reweighting and deterministic reweighting.

Iterative Proportional Fitting (IPF) is according to Moretti et.al. [3] a method for deterministic reweighting to map the distribution of coarse granular data (deaggregated data) to fine granular data (aggregated data). For this a unambiguous mapping between deaggregated and aggregated data is necessary. This mapping does not exist between a 1km grid and administrative boundaries of municipalities, so that IPF alone would not be sufficient to synthesize a reference population.

Basse et.al. [4] deal with exactly this problem using cellular automata (CAs) and neural networks. Cellular automata are used to simulate land use for grid cells in border regions, while neural networks are used as a probabilistic model. However, the application of neural networks requires a large demographic database, which is not available in ECR. Therefore, the combination of IPF and CA in this paper is necessary for population synthesis in order to counteract problems in border regions and insufficient data volumes.

3. Concept

For the concept, the data basis should first be considered. The aggregated data for IPF contain the demographic information with 10-year age classes and without reference to the two gender classes in a 100m grid of the 2011 census. The deaggregated data contain demographic data at municipality level with 1-year age classes with reference to the gender classes and sparsely smoker status or occupation as cancer specifc attributes. Additional data are information from the census on households and families and are used in the integration of CA with IPF. From the deaggregated data, statistical ratios such as the proportion of a 1-year class in a 10-year class are extracted as additional information.

This results in a vector $U = \{U_1, U_2, \ldots\}$ of attributes that are to be generated with a function $P_{U,C}$ by assigning a probability $p_{i,j}$ to each attribute characteristic $u_{i,j} \in U_i$, taking into account the deaggregated data as context C. $P_{U,C}$ gets as argument a vector $A = \{a_1, a_2, \ldots\}$ of the aggregated data so that a random number $\pi \in [0;1]$ together with A calculates a vector $u = \{u_1, u_2, \ldots | u_i \in U_i\}$ representing a single person with all required attributes. A CA then checks the conformity of the generated data with the family, household and demographic data and gradually adapts the demography so that the information on families and households is reflected in the demographic data, but meets the constraints of the municipality demography.

In the application, $U = \{age, gender, residence\}$ is used with 1-year age groups up to 80 and the 100m grid of the census as a place of residence. The vector $A = \{age_{10}, residence, gender_n, households, families\}$ contains the 10-year age groups up to 80 of the census, the same place of residence and the number of men and women living in the grid cell together with the information on families and households for the CA. Other attributes such as smoker status or hormone receptor information for lung cancer or breast cancer, for example, are also possible for U and A if the necessary data are available. Based on epidemiological cancer registration, smoking status information is rare and hormone receptor information is not available, but hormone receptor information should be reported in the clinical cancer registration for estrogen in breast cancer.

After generation, an additional function $\bar{p}$ is applied, which can serve as a bias. $\bar{p}$ is particularly useful if generation is to be carried out for years other than the census year, since population changes and changes in administrative boundaries can be modelled here. On the other hand, this function can be used to correct inconsistencies between different data sources.

4. Implementation

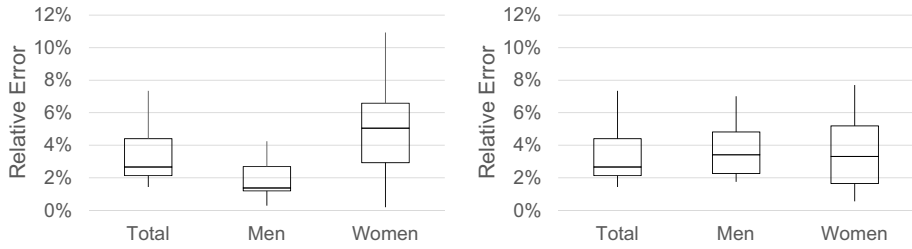
The data of the census are available in tables with about 6 columns, but 61 million entries. Partitioning based on characteristics (age, gender) is essential to shorten access times

to the data. The geo references of the census are used to link the sub-data sets of the aggregated data and a mapping to municipalities is used to link the deaggregated data.

Population synthesis uses IPF for an initial statistical distribution of personal characteristics in the grid cells of the municipality. On the basis of this distribution, CA is used to generate people probabilistically until the statistical distribution is approximated, taking into account household and family data, or a certain number of iterations has been reached. However, the average number of persons per grid cell is 6, so that the empirical law of large numbers cannot be applied to the probabilistic method to approximate the statistical distribution. An adaptation of probability from [5] helps to achieve faster convergence despite small numbers of people. For this purpose, logistic growth is used as an adaptation function with the turning point $(1, p_i)$ for the basic probability of a feature, and the limits are defined by $p_u = 1.5 \cdot p_i$ and $p_l = 0.5 \cdot p_i$. The abscissa x counts the occurrence of an event and the higher x is, the greater the probability of the counter-event, whereby the occurrence of the counter-event x can either be set to 1 (hard reset) or reduced by 1 (soft reset). $\bar{p}$ is trained with municipality data from the census year 2011 and a target year, e.g. 2017, so that the generation can also be used for other years.

5. Evaluation

The evaluation of the data quality on the 1km grid is not possible, because data with comparable accuracy are missing. Instead, the generated 100m grid cells of the map step are aggregated at the municipality level and compared with the previous municipality data for the year 2011. $\bar{p}$ thus has no influence on the data quality, since the identity function is used for 2011. Since the procedure is probabilistic, several iterations should be considered for evaluation. The influence of probability adaptation (PA) should also be measured to determine the necessity and contribution to stability of the adaptation. A total of 12 iterations with and without probability adaptation were performed, with Figure 2 showing the best results of the iterations based on the relative error for the generated gender distribution in the municipality average.



(a) Without probability adaptation.

(b) With probability adaptation.

Figure 2. Relative error in Lower Saxony in 2011 with and without probability adaptation.

The total plots in Figures 2a and 2b are identical and reflect the relative error of the number of persons generated between the 100m grid and the municipality level. The aim of the procedure should be to distribute the error as evenly as possible over the individ-

ual characteristics in order to achieve stable convergence. The generation without PA in Figure 2a shows a clearly higher variance, in particular the even distribution of the error on the characteristics does not take place, because a low error for one gender results in a high error for the other gender. Without PA, for example, the maximum error can vary between 4% and 12%, while with PA it can vary between 7% and 8% as Figure 2b indicates. Overall, PA is necessary as a factor for the stabilization and reliability of the procedure, whereas the synthesized population for Lower Saxony deviates from the actual population by an average of 2.7% and can therefore be used for the analysis of local accumulations.

6. Conclusion

Currently, $\bar{p}$ is used for 2011 as an identity function or to represent the population for other years. It is conceivable that this function can also be trained to minimize errors, so that overall errors and thus also the generation error of the characteristics decrease and can lie within the 99% confidence interval. In addition, the procedure can also be used for regions other than Lower Saxony or for data with other granularities. Furthermore, the method can also be applied for the approximation of health information described in [6], e.g. supply chain company networks on different levels (personnel, specialist groups, company groups) for drug research or for clinical studies, in order to dissolve clusters if necessary and to reduce study costs by data approximation in preliminary studies.

This paper has given an overview of various granularities of demographic data in epidemiological cancer registration and a probabilistic procedure for approximating actual demographics for a 1km grid using a combination of Iterative Proportional Fitting (IPF) and cellular automata (CA). For this purpose, municipality-level data were combined with data from the 2011 census for Lower Saxony, so that a statistical distribution was created by IPF as a framework for the actual generation of the population by CA in a 100m grid. The generation error in relation to known data is on average 3%, but a detailed evaluation on 100m or 1km grids was not possible because no reference data were available.

References

[1] C. Stegmaier, S. Hentschel, F. Hofstädter, A. Katalinic, A. Tillack, and M. Klinkhammer-Schalke, *Das Manual der Krebsregistrierung*. D-82110 Germering/München: W. Zuckschwerdt Verlag GmbH für Medizin u. Naturwissensch., 2018.

[2] D. Ballas, T. Broomhead, and P. M. Jones, "Spatial microsimulation and agent-based modelling," in *The Practice of Spatial Analysis*, pp. 69–84, Springer, 2019.

[3] A. Moretti and A. Whitworth, "Evaluations of small area composite estimators based on the iterative proportional fitting algorithm," *Communications in Statistics-Simulation and Computation*, pp. 1–18, 2019.

[4] R. M. Basse, H. Omrani, O. Charif, P. Gerber, and K. Bódis, "Land use changes modelling using advanced methods: Cellular automata and artificial neural networks. the spatial and explicit representation of land cover dynamics at the cross-border region scale," *Applied Geography*, vol. 53, pp. 160–171, 2014.

[5] J. W. Weibull, "What have we learned from evolutionary game theory so far?," tech. rep., IUI Working Paper, 1997.

[6] O. Tamburis and I. Bonacci, "Clusters and communities: raising the bar towards open innovation 2.0 paradigms," *International Journal of Pharmaceutical and Healthcare Marketing*, vol. 13, no. 3, pp. 288–305, 2019.