
simSALUD: A Web-based Spatial Microsimulation to Model the Health Status for Small Areas Using the Example of Smokers in Austria

Bernhard KOSAR and Melanie TOMINTZ

Carinthia University of Applied Sciences, Villach / Austria · b.kosar@cuas.at

This contribution was double-blind reviewed as full paper.

Abstract

The Austrian Federal Ministry of Health aims to improve the health of all people living in Austria and to decrease social inequalities in health and other areas. This leads to careful planning and distribution of the available health care resources to meet the targets of the government. Health related data mainly exists at federal state level, which provides an interesting overview of the current health situation, but for regional planning authorities, this data is often insufficient as they can hide pockets of high and low health prevalence in certain municipalities. The research project SALUD (SpatiAL microsimUlation for Decision support), funded by the Federal Ministry for Transport, Innovation and Technology, and the Austrian Science Fund, focuses on building a spatial microsimulation model for Austria. This paper presents the use-case of a spatial analysis of estimated smokers for municipalities in the year 2011. This is done by combining survey and census data to model, in particular, the health issues of small areas (e.g. municipalities, districts) based on individuals or households where no data exists, so called missing data. Within this project, a first framework of simSALUD (www.simsalud.org/simsalud) is developed, which is a web-based spatial microsimulation application (designed as wizard) for health decision support, as there is currently no flexible and freely-available web mapping framework in the area of spatial microsimulation. The application is designed for experts as well as for non-experts to simulate their own data that can be visualized afterwards without any programming skills. The results will be valuable in supporting policy decision makers by indicating where spatial health inequalities exist, so that they can distribute their resources more efficiently and thus reduce health inequalities.

1 Introduction

A core aim of the Austrian health care system is to provide equal opportunities for all persons to receive health care, independent of factors such as education or income, so that everyone has the same chance to stay healthy (or get back to full health). However, it is recognized in Austria, as well as many other countries, that the reality is different as health inequalities exist and are increasing as the gap between social classes widens. Nowadays, health problems, especially related to a bad life style, are problematic, e.g. obesity, smoking, diabetes mellitus type 2, but they are preventable by doing more exercise, eating

more healthy and going to medical check-ups. For health care planners, this leads to a careful planning and distribution of available health care resources which are becoming scarce. For effective actions at a regional level, it is important to know where people in need are located, and if there are differences in the spatial distribution of certain diseases or health problems. One problem which we have in Austria is the missing health related data at a local (e.g. district or municipal) level. Such data is mainly provided by Statistics Austria, and usually only exists at the federal state level. Although providing an interesting overview of the current health situation, this data is often insufficient for regional planning. Furthermore, the health survey published by Statistics Austria is conducted at irregular intervals, and the last one was conducted in 2006/07. The aim of our research project SALUD is to estimate health related data for small areas (e.g. municipalities) by applying an approach called spatial microsimulation, where different data sets (spatial and non-spatial data sets) are linked to create synthetic microdata, see Figure 1. This is important to identify *where* health inequalities (e.g. smokers, diabetes patients, etc.) are located (e.g.: municipality or district) in order to support health policy planning. This paper demonstrates a spatial microsimulation approach on simulated smokers at the municipality level in Carinthia. Other similar models focusing on health care demand with different approaches exist (TOMINTZ et al. 2009). Recent OECD statistics show that 23.2% of the adult population smoke daily in Austria, thereby lying 2.07% above the other 41 OECD countries. Furthermore, Austria tended to follow the general pattern marked by high smoking rates among men with 27.3 and women with 19.4% (OECD 2013).

Do regional health inequalities exist?

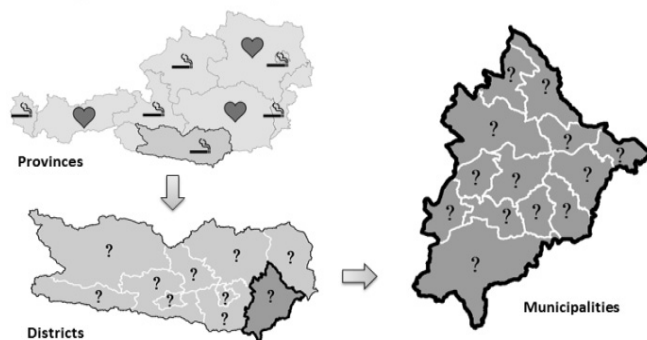


Fig. 1: Usually unavailable health related data for small areas is estimated (in this case, for municipalities)

One powerful advantage of spatial microsimulation is the possibility of modeling various what-if scenarios, which allows policy makers to explore the impact that certain policies would have on certain population groups, for example, in the case of the variable smoker:

- Where to distribute “quit smoking information packages“ in order to reach people most in need?
- Which population groups are most affected by increasing cigarette prices and where do these people live?

2 Methodology

Spatial microsimulation (SMS) models have been built since the 1980s and constitute an increasingly well-established technique, mainly aiming to support policy decisions for various areas, including tax, transport, environment and health. Such models can either be dynamic (behaviour over time), or static (constant behaviour), including different algorithms for estimating spatial microdata for small areas. Basically the approach to estimate spatial microdata combines data from different spatial and aspatial data sources and uses distinct algorithms. The workflow of the model process is shown in Fig. 2.

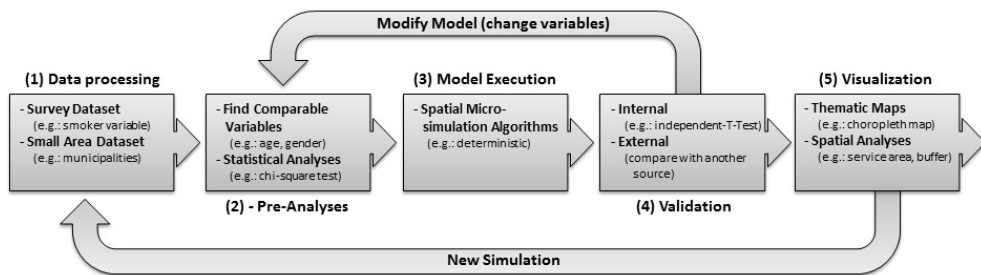


Fig. 2: Workflow of the spatial microsimulation model using the simSALUD framework

The workflow, described in the next subchapters, can be divided into five parts for building a spatial microsimulation. These comprise the data processing, pre-analyses, model execution, validation and the visualization/spatial analyses of the estimated results.

2.1 Data processing

Health related data usually only exist at a federal state level, however, for regional planning these data are not suitable due to their coarse spatial resolution. This survey dataset is primarily provided by Statistics Austria. The institution conducts health related surveys at irregular times, and the last extensive one at a national level was conducted in 2006/07 with a sample size of 15.474 persons and 658 variables. The survey data on demographics, socio-economic factors, and health issues, is available for people aged 15 and above. One variable includes a weighting for each person, to compensate for the bias of the survey that corresponds to the total population of all Austrian citizens who are older than 15 years of age. The non-health related spatial census dataset from 2011 is also provided by Statistics Austria and consists of 39 variables that are digitally collected by obtaining data from the registers office.

2.2 Pre-Analyses

The applied static deterministic reweighting approach requires linking both datasets based on common variables, e.g. age and sex. These constraints should also be strong predictors (i.e. characteristics of a person) of the variable being estimated (e.g. smokers). This process is essential for an accurate and robust model. The datasets need to be analyzed for suitability and processed in a form that is required for the simulation. Furthermore,

statistical tests (e.g. chi-square, binary logistic regression) and intensive literature research are performed on the dataset, to explore which variables are most suitable as constraint variables for the model. Taking into account of all these preconditions (common variables, statistical tests), 5 variables (27 categories), listed in Tab. 1, are chosen and remain at municipality level.

Tab. 1: Constraint variables selected for the spatial simulation model

Variable	Categories
Household size	Non-private household, 1 person, 2 persons, 3-persons, 6 above
Sex	Male, female
Age	15 - 19, 20 - 24, 25 - 29, 30 - 34, 35 - 39, 40 - 44, 45 - 49, 50 - 54, 55 - 59, 60 - 64, 65 above
Marital status	Single, married, widowed, divorced
Occupational status	Employee, employer, self-employed person, family workers, non-employed person

In contrast, the variables nationality and highest school-leaving qualification are only weak predictors (e.g. no statistically significant association). In a next step, the categories of the constraint variables are adjusted and recoded according to the constraint categories of the census data. If for example the constraint age is categorized in the survey dataset in 2 groups (15-50 years and 51-90 years), and in the small area dataset in 3 groups (15-50 years, 51-70 and 71-90 years), an adjustment has to be made (summarize the small area groups into 2 groups) to get the same comparable categories.

2.3 Model Execution

Based on Edwards & Clarke's (2012) combinatorial optimization, the simulation uses a static deterministic reweighting approach by combining national survey data (Austrian Health Survey 2006/07) and census data from 2011. In general, deterministic reweighting as a spatial microsimulation technique uses algorithms where common variables in both datasets are used to iteratively calculate the probability of an individual to "live" in a certain area. At first, each person starts with a probability of 1 to live in each area, and as the model cycles through the constraints this probability is adjusted given each area's population profile (SMITH et al. 2011). These new weights for each individual (probability to live in this area) with their characteristics are then used to estimate health related data for small areas. Within the simSALUD framework, the prepared .csv files can be uploaded, and the simulation steps are guided through a wizard. After the first algorithm an additional algorithm can be used (tick checkbox) to allocate only whole people to the final dataset. This iterative optimization methodology (BALLAS et al. 2005b) converts the values of the first output from decimals to integers.

2.4 Validation

After running the model, the results have to be validated to verify that the model estimates are robust, ideally by using both external and internal validation. External validation requires additional comparable external sources which do often not exist or are not available due to confidential reasons. For the internal validation the simulated data for all five common variables (constraints) need to be compared with the existing census data, as the census data holds complete information about each person for the simulated area. Several validation methods were applied, including standardized absolute error, and independent-samples T-Test (see chapter 4 Results). It would be expected – and has been approved – that there is no significant difference between the two datasets, which is an indicator that the model estimates are robust. This allows for the conclusion that if the common constraint variables fit well, it can be assumed that also the simulated health variable (e.g. smokers) estimates fit well and are close to reality.

2.5 Visualization

The outputs of the simulation can be either visualized in form of a map on the simSALUD web-based framework, or can be exported after the model run for further analyses in common geographic information software products. The map results allow decision makers to show pockets of areas of estimated regional health inequalities (e.g. high/low smoking rates) in municipalities.

3 Implementation

At the moment there is no public web-based spatial microsimulation application available which guides people to build their own models. They need to customize the models to their needs, which is time and cost consuming. Current simulations with common approaches (e.g. Microsoft Excel, Access, Java applications, etc.) are either not user friendly or simply not realizable for such big data, so that alternatives had to be found. Some microsimulation applications exist on the market (ABELLO & BROWN 2007, DE MENTEN et al. 2012, MORRISSEY et al. 2013 & HARLAND 2013). Many applications are not flexible enough regarding user handling (e.g.: changing model input parameters, etc.) and are not accessible for everyone, or do not support visualization of the results. A solution is simSALUD, a web-based spatial microsimulation framework in Java, where first the deterministic reweighting approach is implemented to simulate the smokers at municipality level. For policy decision makers, it is of high interest to visualize the created small area data in form of a map, and/or link it with additional data for location analysis. For that purpose, one scope of the project was to develop a flexible web-based spatial microsimulation for health decision support, with an easy-to-use graphical user interface (wizard) for experts and non-experts (especially health policy makers), with visualization and simple analysis functions of the microdata feasible without programming skills. This user-friendly web simulation is based on Apache's open source web application framework Struts2 and uses ESRI's JavaScript API and ArcGIS for Server for the map visualization. ESRI's JavaScript API, included with ArcGIS for Server, is based on the GeoService Representational State Transfer (REST). The GeoService REST Specification provides a standard way for web

clients to request maps, features, attributes, and image information from a GIS Server (ESRI 2010). The framework consists of three modules (see Fig. 3) called (i) Spatial Microsimulation Framework, (ii) User Interaction Wizard and the (iii) Web Mapping Framework.

- The User Interaction Wizard (UIW) represents the graphical user interface of the web simulation. The wizard allows the user to upload all required datasets (spatial and aspatial) into a database, as well to match the survey and the census data using a drag and drop technique. The UIW also contains the results manager, where the user can download the simulated data file and see all the results visualized. The visualization page presents an interactive map with different standard functionalities.
- The Spatial Microsimulation Framework (SMF) can also be seen as the server module. The SMF represents the calculation unit of the system. On the server the algorithms are applied, so that for the client system (e.g.: notebook, etc.) no further calculation performance is necessary.
- The third Web Mapping Framework component can be seen as geoprocessing tools, and will enable individual spatial analysis to be performed on the data, to answer various what-if scenarios. The Map Provider (ArcGIS for Server) with its Map Communicator will handle the user requests, process them, and send the finished product back to the client. For this purpose the web clients uses ArcGIS's geoprocessing services which contain certain geoprocessing tasks (e.g. buffer, service area). The service captures the data, processes it, and returns the output in form of maps.

simSALUD

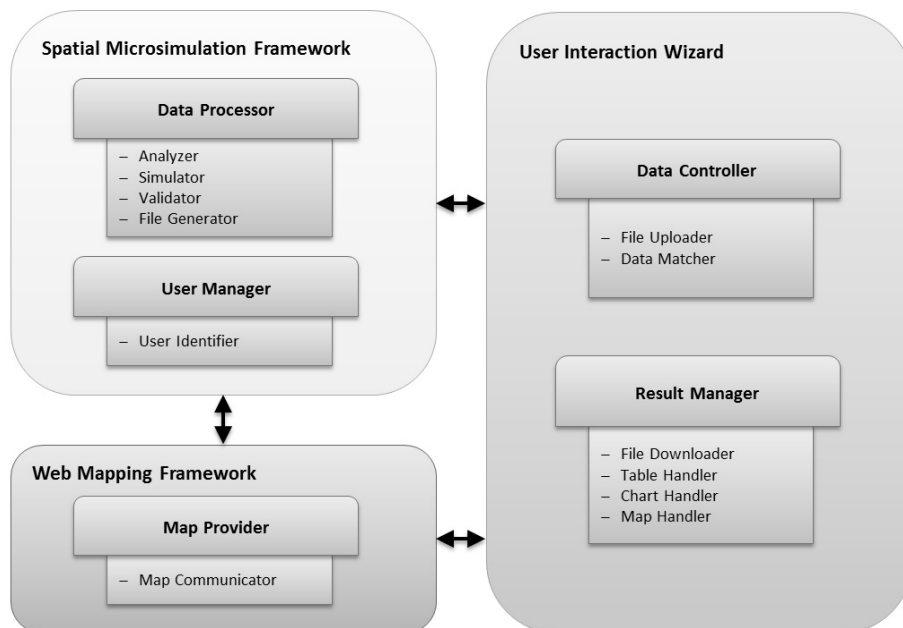


Fig. 3: System architecture of the simSALUD Framework

The website also supports responsive design, which dynamically adjusts the layout of the page depending on the device used (mobile device, tablets or PC). The current beta version of the web application is already evaluated regarding usability, with methods such as think aloud protocol (LEWIS 1982) and Nielsen's Usability Heuristics (NIELSEN 1994).

4 Results

The first case study focuses on simulating smokers in Carinthia since Austria has a high smoking population, especially amongst adolescents. The observed survey dataset from 2006/07 includes 15474 participants with several variables. These constraints must be strong predictors of the simulated variable and comparable with the census dataset from 2011. The projected average rate of current smokers (15 years and over) in Austria is 23.2%. The stepwise logistic regression analysis as well as the chi-square test showed that the variables that should be used as constraints are sex, marital status, age, household size and occupational status. Other variables such as nationality or birthplace showed no significant improvement in the model's fit. After the model run, the validated results showed that the small-area estimates for the model were robust with no significant differences between the simulated and actual dataset. The linear regression analyses indicated a very high coefficient of determination for all variables except the variable employer. The Percentage of the Standardized Absolute Error (PSAE) and an equal variance t-test were also applied for all variables. This is summarized in Tab. 2.

Tab. 2: Summary of the validated model outputs

Variable	Category	PSAE (%)	P value	Variable	Category	PSAE (%)	P value
Sex	Male	4.57	0.74		55 - 59	2.04	0.30
	Female	4.69	0.75		60 - 64	1.88	0.28
Marital status	Single	2.14	0.92		65 above	7.67	0.18
	Married	3.20	0.97	Household size	Non-private household	0.18	0.71
	Widowed	1.70	0.48		1 person	3.63	0.62
	Divorced	0.91	0.87		2 persons	2.70	0.74
Age	15 - 19	3.40	0.11		3 persons	1.70	0.94
	20 - 24	3.89	0.13		6 above	0.50	0.77
	25 - 29	3.23	0.27	Occupational status	employee	3.95	0.88
	30 - 34	1.37	0.55		employer	6.39	< .001
	35 - 39	0.85	0.70		self-employed person	0.29	0.70
	40 - 44	0.68	0.92		family workers	0.12	0.14
	45 - 49	0.91	0.90		Non-employed person	7.54	0.59
	50 - 54	0.88	0.77				

The PSAE shows the highest score, with a total of 7.67% under- or over-estimation values to the actual dataset in the category "65 above". The results of the t-test analysis also demonstrated no significant differences between the simulated and actual dataset, except – again – for employer. The simulated results in Fig. 4 show a spatial variation of smoking rates at the municipally level. Especially municipalities in the districts of Spittal an der Drau, Villach Land, Klagenfurt Land and Völkermarkt have a significant variation of low and high percentages of smokers.

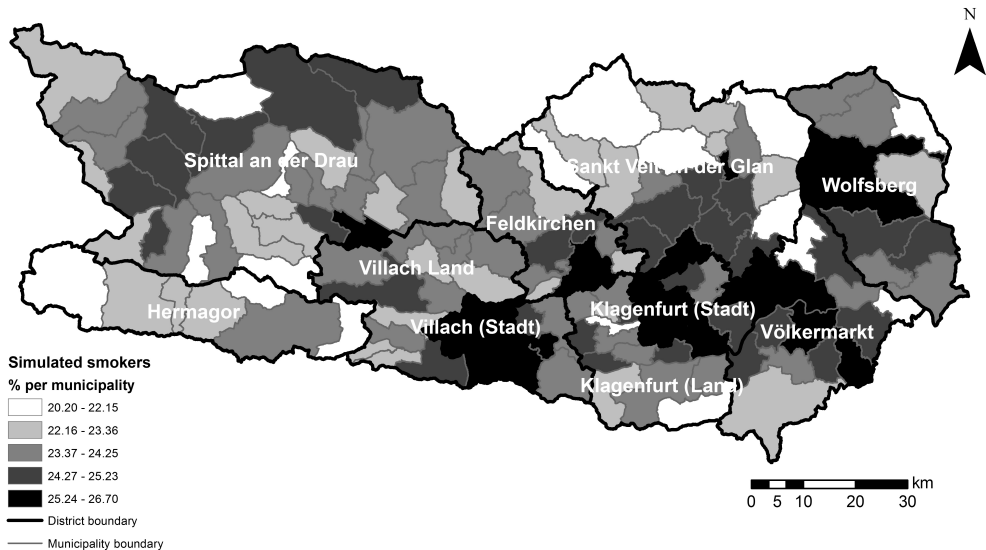


Fig. 4: The figure shows the simulated smoking population in Carinthia in 2011. Regional inequalities can be identified, especially in the districts of Spittal an der Drau and Sankt Veit an der Glan.

The output will be presented to Carinthian health care planners to discuss possible social influences and where actions are required. The first prototype of simSALUD was also introduced on a conference to our international project partner, the University of Leeds - who are leading experts in this area - and this approach enjoyed high popularity. The main advantage of the web-based application is that it can be used without programming skills to perform simulations or to analyze the simulated data on maps in order to identify hotspots more easily. Based on these results it will be interesting to model what-if scenarios, e.g. where do smokers aged 25 and 39 live, to simulate effects of possible governmental actions on certain population groups.

5 Discussion

The first algorithm which is implemented in the simSALUD framework is the static deterministic reweighting method. However, current project results demonstrated that this algorithm is strongly impacted (very high variances) by the choice and number of the constraint variables. Therefore, the input constraints have to be good predictors of the simulated variables and should be highly correlated with the output variables. But the low number of available census constraints in Austria is strongly limited, and this restricts the number of strong predictor variables in Austria. Other algorithms (HARLAND et al. 2012) may achieve better results that will be tested in the next months. High costs can be considered as another negative aspect regarding the use of appropriate census data. In Austria the data are provided in different packages and the higher the spatial resolution the higher the costs depending on the area, even when they are needed for academic/research purposes. It should also be considered that the time of data collection does not match

between the survey data from 2006/07, which is only collected at irregular intervals, and the census data from 2011. This gap will only be closed when the next Austrian health survey is published in 2014/2015. The results of this next survey will be very interesting in terms of analyzing trends such as, for example, how the smoking behavior has changed. However, the collection of detailed health related data only takes place approximately every 10 years, yet again resulting in disparities amongst the different data sources. Indeed other health related data exists in Austria, but it is not representative (small sample size).

6 Conclusion and Outlook

Spatial microsimulation models for health decision-making can make a major contribution to policy analysis as these models are not widely established and distributed yet, especially not in Austria. This conference paper shows the first results of applying a spatial microsimulation model to simulate the smoking population at municipality level in Carinthia, Austria, as no data on smoking exists at this geographical scale. For that purpose, within the framework simSALUD, a deterministic reweighting approach is implemented, constituting the first attempt of a web-based spatial microsimulation for decision support for experts and novices. The simulation allows different topics to be addressed which are high on the Austrian policy agenda, including smoking, obesity, nutrition, fertility or cancer related diseases. This Web application is in continuous development and improvement throughout the research project. The current simulation is based on the Austrian Health Survey 2006/07 and the census data from 2011. These results are interesting as local variations within municipality level are found. Especially interesting are the districts Spittal an der Drau, Klagenfurt Land, Sankt Veit an der Glan and Völkermarkt, as the percentages of the smoking population ranges from lowest to highest within these districts. The district Hermagor in the south-west of Carinthia, on the other hand, has the lowest smoking population of all municipalities, whereas especially the capital cities in the south-east show the highest rates. The results are validated using several methods such as the Standardized Absolute Error, to explore which categories of the constraint variables have a good fit and which are over- and under predicted by comparing the constraint variables of the census with the simulated constraints. Additional validation methods such as regression analysis or the equal variance t-test were undertaken to calculate the correlation of the model datasets. The results of the simulation should illustrate the importance of small area health data and the advantage of the small area modelling approach to governmental health departments. Further investigations will explore possible reasons for higher and lower smoking rates for certain municipalities with expert knowledge. In addition it is aimed to define what-if scenarios for future modelling, so that the Carinthian health decision makers can benefit from the results. Another outlook is to model the smoking population for the whole of Austria at the municipality level. The simulation should support decision makers for future activities, to decrease regional health inequalities.

Acknowledgments

This research project SALUD is funded by the Federal Ministry for Transport, Innovation and Technology (bmvit) and the Austrian Science Fund (FWF) (project number TRP280-

G16). The census data is provided by “STATcube – Statistical Database of STATISTICS AUSTRIA”.

References

- BALLAS, D., CLARKE, G., DORLING, D., EYRE, H., THOMAS, B. & ROSSITER, D. (2005a), SimBritain: a spatial microsimulation approach to population dynamics. *Population Space and Place*, 11, 13-34.
- BALLAS, D., ROSSITER, D., TOMAS, B., CLARKE, G. P. & DORLING, D. (2005b), *Geography matters: Simulating the local impacts of national social policies*. York, Joseph Rowntree Foundation.
- DE MENTEN, P., DEKKERS, D., BRYON, R., LIÉGEOIS, P., WAGENER, R. & O'DONOGHUE, C. (2012), LIAM 2: a new open source development tool for the development of discrete-time dynamic microsimulation model. Paper presented at the séminaire scientifique international, Caisse des Dépôts, Bordeaux, France, 15 November, 2012.
- EDWARDS, K. L. & CLARKE, G. (2012), SimObesity: Combinatorial Optimisation (deterministic) model. In: TANTON R. & EDWARDS, K. L. (Eds), *Spatial Microsimulation: A Reference Guide for users*. Springer, SBM, NL.
- ESRI (2010), *GeoServices REST Specification Version 1.0 [White paper]*. <http://www.esri.com/library/whitepapers/pdfs/geoservices-rest-spec.pdf>.
- HARLAND, K. (2013), *Microsimulation Model User Guide (Flexible Modelling Framework)*, School of Geography, University of Leeds, Leeds, LS2 9JT, United Kingdom.
- HARLAND, K., HEPPENSTALL, A. J., SMITH, D. M. & BIRKIN, M. (2012), Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies & Social Simulation*, 15 (1).
- LEWIS, C. H. (1982), Using the “Thinking Aloud” Method In Cognitive Interface Design (Technical report). IBM T.J. Watson Research Center RC-9265. Yorkworn Heights, NY, T. J. Watson Research Center
- NIELSEN, J. (1994), Heuristic evaluation. In: NIELSEN, J. & MACK, R. L. (Eds.), *Usability Inspection Methods*. John Wiley & Sons, New York, NY.
- OECD (2013), Smoking. In: *OECD Factbook 2013: Economic, Environmental and Social Statistics*, OECD Publishing. <http://dx.doi.org/10.1787/factbook-2013-98-en>.
- SMITH, D. M., PEARCE, J. R., & HARLAND, K. (2011), Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. *Health & Place*, 17, 618-624.
- TOMINTZ, M., CLARKE, G. & RIGBY, J. (2009), Planning the location of stop smoking services at the local level: a geographic analysis. *Smoking Cessation Studies*, 4 (2), 61-73.