Health & Place 26 (2014) 47–52

Contents lists available at ScienceDirect

Health & Place

journal homepage: www.elsevier.com/locate/healthplace

Small area estimation of obesity prevalence and dietary patterns: A model applied to Rio de Janeiro city, Brazil

Guido Cataife*

IMPAQ International, 10420 Little Patuxent Pkwy, Suite 300, Columbia, MD 21044, USA

ARTICLE INFO

Article history:
Received 24 May 2013
Received in revised form
23 October 2013
Accepted 3 December 2013
Available online 11 December 2013

Keywords:
Obesity
Overweight
Small area estimation
Spatial microsimulation
Brazil

ABSTRACT

We propose the use of previously developed small area estimation techniques to monitor obesity and dietary habits in developing countries and apply the model to Rio de Janeiro city. We estimate obesity prevalence rates at the Census Tract through a combinatorial optimization spatial microsimulation model that matches body mass index and socio-demographic data in Brazil's 2008–9 family expenditure survey with Census 2010 socio-demographic data. Obesity ranges from 8% to 25% in most areas and affects the poor almost as much as the rich. Male and female obesity rates are uncorrelated at the small area level. The model is an effective tool to understand the complexity of the problem and to aid in policy design.

1. Background

The shift towards high-energy-dense diet and a sedentary life style has turned obesity from a developed-world phenomenon into a global epidemic (Caballero, 2007; James et al., 2001; Kelly et al., 2008). Because obese populations are at a higher risk of developing non-communicable diseases, this phenomenon requires immediate attention.

The World Health Organization recommends prevention strategies, such as public education campaigns, to fight obesity (WHO, 2000). Since the success of this intervention depends on resources not being wasted (WHO, 2000), targeting specific groups becomes critical (McLaren, 2007). However, modern monitoring and surveillance systems needed to identify the target population are seldom available in developing countries (Prentice, 2006). This paper proposes a practical solution to this problem and presents the case study of the Brazilian city of Rio de Janeiro (RJ).

Obesity prevalence rates, as well as food intake habits, vary by demographic and socioeconomic groups (Clarke et al., 2009; McLaren, 2007; Monteiro et al., 2004; Sobal and Stunkard, 1989). Since socio-economic and demographic characteristics have a strong spatial component at the small-area level, it is plausible to target different socioeconomic and demographic groups by targeting different small areas. Ideally, the strategy would rely on Census Tract level data. A Census Tract represents the smallest geographical area at which census data are available—on average, a Census Tract in RJ comprises as few as 618 individuals.

We use small area estimation (SAE) to produce measures, graphs, and maps to show how obesity and overweight prevalence can be monitored with high granularity at very low cost. SAE is a group of techniques used to fill the data gap over small geographic areas (Ballas et al., 2005b; Rao, 2003; Tanton and Edwards, 2013). Although SAE applications are computationally intensive, their associated economic cost is negligible since they do not require major data collection efforts. Other appealing features of SAE are its wide scope of applicability and the ease with which it can be integrated into complementary methodologies, such as geographic information systems (GIS) and practically any analytical technique (e.g., regression analysis). For all these reasons, SAE applications are being increasingly used for policy making. Prominent examples are the Small Area Income and Poverty Estimates (SAIPE), the Small Area Health Insurance Estimate (SAHIE) carried out by the U. S. Census Bureau (Huang and Bell, 2004; O'Hara, 2008), and the SAE of obesity and healthy diet carried out by the National Observatory of Obesity in the UK (NOO, 2013). The relative novelty of the approach, along with the lack of a unified theoretical framework, may be why developing countries, which usually have tighter budgetary constraints and more deficient information systems, have not yet exploited SAE.

We show the utility of SAE for monitoring and surveillance of obesity and overweight by discussing a number of prevalence estimates and food intake applications pertaining to RJ city. We also briefly discuss underweight prevalence to test the related “dual burden of disease” hypothesis, which says that households

* Tel.: +1 443 367 0088.
E-mail address: gcataife@impaqint.com

1353-8292/$ - see front matter
http://dx.doi.org/10.1016/j.healthplace.2013.12.004

with both obese and undernourished members are increasingly observed in developing countries, mainly as the result of changes in dietary patterns (Caballero, 2005).

2. Data and methods

Several approaches have been developed to generate SAE (Ghosh and Rao, 1994; Pfefferman, 2002; Rao, 2003). We use spatial microsimulation (Ballas et al., 2005a, 2005b; Simpson and Tranmer, 2005) based on combinatorial optimization (CO) (Voas and Williamson, 2000). Publicly available software has the capacity to produce this particular type of SAE estimates (Williamson, 2007b). The software has been used previously (Hermes and Poulsen, 2012).

The approach consists of the optimal selection—from a survey conducted in a large area—of records for persons and households that are known to inhabit the targeted small area. Two sources of data are required. The first is a survey conducted in a "large area," which must satisfy the following properties: (i) it must include the targeted variables (the variables that one is seeking to obtain at the small area level); (ii) it must include socioeconomic and demographic variables that correlate with the targeted variables; and (iii) these socio-economic and demographic variables must be available at the small area level from census tables. The second source of data is census tables at the small area level. Census data are used to determine the number of persons and households in the targeted small area and to provide the tables on the demographic and socioeconomic characteristics of the small area population. In essence, the socioeconomic and demographic variables available in both the large area survey and the census are used to identify the large area survey records that best fit the characteristics of the small area population. A "simulated" small area survey consists of the subset of records from the large area survey that best matches the socioeconomic and demographic characteristics of the targeted small area. Further discussion of the CO microsimulation methodology is available elsewhere (Williamson, 2013).

We simulate indicators of obesity, overweight, and underweight at the individual level for adults (20 years of age or older) and households' sweet soda consumption (liters) data to produce datasets with these same variables but at the Census Tract level. The Body Mass Index (BMI) formula (BMI=weight(kg)/[height (m)]^2) is used to determine whether each individual is overweight (BMI between 25.0 and 29.9), obese (BMI $>$ 30.0), or underweight (BMI $<$ 18.5) (WHO, 2013).The obesity and overweight estimates are directly relevant to assessing the magnitude of the body mass excess problem. Sweet soda consumption estimates are produced since recent hypotheses suggest that high-energy-dense diets, such as those consumed in the U.S. and other developed countries, characterized by a large intake of sweet soda (Caballero, 2007), are one of the factors behind the global obesity epidemic (James et al., 2001; WHO, 2000). Finally, the underweight variable allows us to test the "dual burden of disease" hypothesis, which says that households with both obese and undernourished members are increasingly observed in developing countries, mainly as the result of changes in dietary patterns (Caballero, 2005).

We further explore geographical variation by contrasting two RJ neighborhoods that have very different sociodemographic characteristics: the bohemian and affluent *Humaita* and the low-income community *Cidade de Deus*. The specific small area targeted can affect the goodness of fit of the model: "typical" areas are likely to result in a better fit than their counterparts since they are bound to be better represented in the large area survey (Smith et al., 2011). Given their atypical socio-demographic profiles, our target areas are challenging cases for SAE. Our focus on RJ city relies on its demographic and economic relevance, but it is otherwise arbitrary. The large area survey utilized, the *Pesquisa Orçamentos Familiares 2008–9* (POF 2008–9) (IBGE, 2011), is a nationally representative family expenditure survey conducted in 2008–2009, with a sample size of 65,000 urban and rural Brazilian households. The survey, designed to capture the household spending patterns during a 1-year period, contains the height and weight data required to calculate the body mass index. Although other surveys also provide these data, we use POF because it allows additional applications and collects information on many socio-demographic characteristics of household members and their living conditions necessary to match the large area survey with census tables. Given the large sample size and wide scope of the data collected, the survey is especially suited to studying food consumption habits (IBGE, 2011). The technique can also be applied to any other nationally representative survey, such as the *Pesquisa Nacional por Amostra de Domicilios* (PNAD), for which scholars are interested in obtaining small area estimates (Viacava et al., 2006). To improve the fit of the model, we use only POF data for households in Rio de Janeiro state, the home of Rio de Janeiro city. Our analytical sample comprises a total of 1932 households with 5695 individuals.

We choose the most highly disaggregated level at which census data are available—the Census Tract. There are over 10,000 Census Tracts in the city, each typically comprising a few city blocks. Eight census tables at the census-tract level were used to generate six person-level and forty-five household-level variables for the matching. Table 1 describes all variables and indicates the source and analytic role. The table also provides summary statistics for the POF 2008–9 variables. Census variables were created based on official IBGE Census 2010 tables (IBGE, 2012a).

The validity of the approach relies on the premise that the variables used to match the survey data with the census tables are significantly correlated with the simulated variables (Edwards and Clarke, 2013; Williamson, 2013). Matching variables typically include age, sex, occupational status, education level, and household income. All the technique requires is correlation between the predictors and the variable to be simulated. Multicollinearity among the predictors is not a concern. For instance, although education level and household income are often highly collinear, they can both be used in the matching process since the ultimate purpose is not to identify causal effects but to maximize predictive power. We use a saturated regression—i.e., a regression that includes all possible interactions between the covariates in the right-hand side of the equation—to assess the maximum attainable predictive power of a microsimulation based on the variables in Table 1. We use a logistic model to regress overweight/obese status (yes/no) and underweight status (yes/no), separately, on the individual term and interaction effects implied by the variables in Table 1. We use *desmat* as a prefix command to run the saturated regression in STATA. Desmat converts each variable into a set of indicator variables (one indicator for each category of the variable) and allows the specification of high-order interaction effects in the regression. The regression uses approximately 5800 RJ state person-level records from POF 2008–9. Although the fully saturated model could not be estimated due to computational limitations, 260 level and interaction effects were included. The fit of the logistic model is evaluated in terms of Nagelkerke r-square (0.27 and 0.52 for the overweight or obese model and the underweight model, respectively), joint statistical significance of all independent variables ($p < 0.001$ for both models), and correct classification of cases (66.8% and 88.8%, respectively). However, as Table 1 describes, in our simulation we have included no interactions between variables. Hence, these statistics should be interpreted as an assessment of the maximum attainable predictive power rather than the actual power achieved by our parsimonious application.