Working Paper Series, N. 10, July 2008

Perceived neighbourhood quality and adult health status: new statistical advices useful to answer old questions?

Pierantonio Bellini, Daniela Lo Castro and Francesco Pauli
Dipartimento di Scienze Statistiche
Università di Padova

Abstract:

Background. Interest in the quantitative effects of neighbourhood characteristics on adult health has recently increased in social epidemiology. Particularly, investigations concern the statistical influence on health of several individual demographic and socioeconomic characteristics and of neighbourhood characteristics as perceived by respondents. We analyze these issues within an original conceptual framework and employing statistical models unusual in this context.

Methods. We use data collected in the Los Angeles Family and Neighbourhood Survey (L.A.FANS) to model the number of hospital admissions occurred to each individual as a function of some individual and neighbourhood characteristics, the latter being related to the individual perceptions about the neighbourhood he lives in. We employ generalized additive models with different distributional assumptions: Poisson, Negative Biomial and Zero Inflated Poisson (ZIP). Such models allow us to estimate (through spline functions) potential non linear effects of the covariates on the response. Moreover, non standard representations are used to overcome difficulties in interpreting the results for ZIP models.

Results. It turns out that perceived neighbourhood characteristics, and in particular the perception of social cohesion, have a significant effect after controlling for individual characteristics relevant to hospital admissions frequency. From a modeling point of view ZIP and Negative binomial models prove to be superior to standard Poisson model.

Conclusions. We have confirmed the role of the neighbourhood where an individual lives in determining his health status. A strength of this analysis is that, due to the choice of the neighbourhood characteristics to be included in the model, the results do not depend of a particular definition of neighbourhood (which is traditionally based on administrative boundaries), since each individual refers his perceptions to his personal definition of it.

Keywords: Neighbourhood characteristics, Health status, Zero Inflated Poisson, GAMLSS

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Francesco Pauli
tel: +39 049 827 4127
francesco.pauli@unipd.it
http://homes.stat.unipd.it/fpauli

Perceived neighbourhood quality and adult health status: new statistical advices useful to answer old questions?

Pierantonio Bellini, Daniela Lo Castro and Francesco Pauli
Dipartimento di Scienze Statistiche
Università di Padova

Abstract:

Background. Interest in the quantitative effects of neighbourhood characteristics on adult health has recently increased in social epidemiology. Particularly, investigations concern the statistical influence on health of several individual demographic and socioeconomic characteristics and of neighbourhood characteristics as perceived by respondents. We analyze these issues within an original conceptual framework and employing statistical models unusual in this context.

Methods. We use data collected in the Los Angeles Family and Neighbourhood Survey (L.A.FANS) to model the number of hospital admissions occurred to each individual as a function of some individual and neighbourhood characteristics, the latter being related to the individual perceptions about the neighbourhood he lives in. We employ generalized additive models with different distributional assumptions: Poisson, Negative Biomial and Zero Inflated Poisson (ZIP). Such models allow us to estimate (through spline functions) potential non linear effects of the covariates on the response. Moreover, non standard representations are used to overcome difficulties in interpreting the results for ZIP models.

Results. It turns out that perceived neighbourhood characteristics, and in particular the perception of social cohesion, have a significant effect after controlling for individual characteristics relevant to hospital admissions frequency. From a modeling point of view ZIP and Negative binomial models prove to be superior to standard Poisson model.

Conclusions. We have confirmed the role of the neighbourhood where an individual lives in determining his health status. A strength of this analysis is that, due to the choice of the neighbourhood characteristics to be included in the model, the results do not depend of a particular definition of neighbourhood (which is traditionally based on administrative boundaries), since each individual refers his perceptions to his personal definition of it.

Keywords: Neighbourhood characteristics, Health status, Zero Inflated Poisson, GAMLSS

1 Introduction

Interest in the effects of neighbourhood or local area characteristics on health status and outcomes has increased in recent years: the feeling is that the context in which people live, as well as personal characteristics, affects their well-being and quality of life. Identifying which neighbourhood qualities and characteristics are more important to health is then a central issue not only for a better understanding of the connection between health and place, but also to assess health status of communities and to inform future health intervention

strategies (Berkman and Kawachi (2000)).

Our work aims to face statistical questions that, in our opinion, need new suitable answers concerning the new (and old) issues concerning the comprehension of phenomena in order to understand the improvement of the quality of life in cities (i.e. the effect or the outcome) by promoting a continuous improvement process of the surrounding physical and social environment – the potential causes or determinants. For instance this is the case of the Healthy Cities Project (HCP, http://www.euro.who.int/healthy-cities) in particular in its IVth phase (2003-2008).

This raises a number of issues, among which there is a technical task, essential in order to obtain HCP (present and future) success: there is in fact a need for a conceptual framework (CF) – a systemic vision emphasizing interaction integration and coherence among different entities to be investigated (Bellini et al. (2002), Murray et al. (2003), Rodella et al. (2003), Bellini and Rodella (2004), McDowell et al. (2004)) serving the purpose of better identifying the scope of interest – and for quantitative tools suitable for measuring the "health" of the city and to discover, even by means of statistical techniques, the links between outcomes and known/unknown (potential) determinants exploiting spatial and temporal dimensions of the phenomenon.

1.1 Conceptual framework (CF)

In order to tackle the task of empirically verifying the influence of neighbourhood on health one obviously needs data but one also needs, in our opinion, a CF, to heuristically link the main quantities involved, and a statistical model, coherent with the conceptual framework, to quantify the relationships among the relevant quantities.

In the field of community health assessment it is not frequent to make explicit reference to an underlying CF, rather, lists of health indicators of various lengths have been proposed: from those traditional in epidemiology (Hancock and Duhl (1998); National Centre for Health Statistics (2003)) to those concerning environmental, social, behavioural and economic aspects (Webster (1996); Hancock et al. (1999); Canadian Institute for Health Information (1999)).

We show our attempt to an original synthesis of the CFs found in the literature (Starfield (2002); ECHI (2004); WHO Commission on Social Determinants of Health (2005, 2007)) in Figure 1. For the sake of conciseness, we depict categories of determinants rather than specific variables, whose examples can be found in the discussion of results. The vertical dimension defines the determinants to outcomes structure, potential mediators (Li et al. (2007)) are depicted in between outcome and determinants; the horizontal dimension represents the population/ecological dimension: from individual to community characteristics with an intermediate class comprising those community characteristics which are defined aggregating the individual ones (it is worth noting that almost for all individual characteristics community level aggregation may be deemed relevant). Links among community characteristics and aggregated individual characteristics are outlined using dotted connectors, while solid connectors depict (plausible) mechanism through which policies spread their effects. It is beyond our intentions to build a model involving all the quantities and relationships depicted in the scheme. Rather, we will consider partial models involving only some of the elements in the CF. It is nonetheless worth embedding such a model in our more comprehensive CF, doing so may help, for example, in combining results from

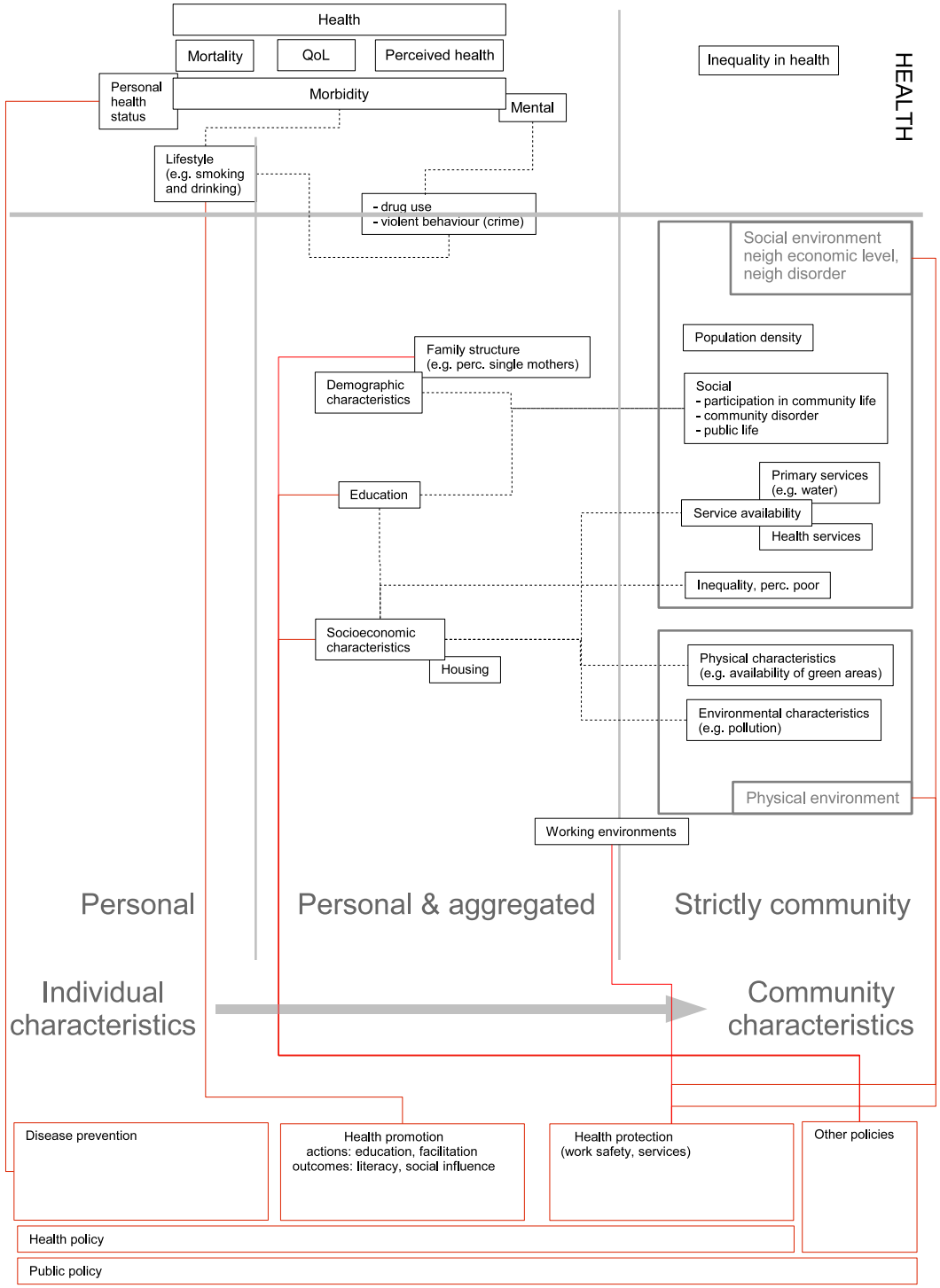

Figure 1: Conceptual scheme. Factors are classified with respect to two criteria. Horizontal classification is from individual characteristic (left) to community characteristics (right); three major classes are recognized: individual factors (those that are meaningful to consider only at an individual level); aggregated individual factors (which may be considered at an individual level but also at a community level by simple aggregation); community factors (which have no correspondent individual level). Vertical classification involves determinants (bottom) to outcomes (top), policies may be seen as exogenous variables.

different studies devoted to different aspects of the issue. Also, one should be aware of the existing links between neighbourhood and individual variables since, to properly estimate effects of neighbourhood variables, one should include in the model all associated individual variables.

Once a CF is developed, the next priority is to verify which of the potential determinants envisaged in the CF have actually an effect on the health status of the population. This kind of analysis are relatively common for some quantities, such as pollutant concentrations, other determinants have been explored less frequently. Neighbourhood quality, which we deem interesting to investigate on, is among the latter group.

1.2 Statistical aspects of the CF

Data and models will be needed to make the CF concrete. Generally speaking, one needs data on health status and other relevant characteristics of individuals. In principle, either administrative or survey data may be used. We do not discuss the advantages and disadvantages of the two kind of sources (about this topic see, for instance, Brackstone (1987) and others). In particular, we consider individual data collected within the Los Angeles Family and Neighbourhood Survey (L.A.FANS, Peterson et al. (2004)), whose main features are briefly discussed in section 2.

As seen in the CF in Figure 1, the boundary between outcomes and determinants is fuzzy: what is an outcome depends on the aim of the analysis. For example, smoking and drinking habits may legitimately be seen as an outcome and a determinant, where in the latter case they may also act as mediators – as defined in (Li et al. (2007)) – between social status and health. Also the fact of suffering from a specific pathology (e.g. high blood pressure) may be seen as an outcome if we study quality of life (high blood pressure directly affects the quality of life) or an intermediate factor if we study mortality. General or specific mortality or morbidity and perceived health are other examples of typical "outcomes".

We choose to consider the number of hospitalizations (in two years) as the health outcome (response variable in our statistical models). Its relationship with neighbourhood quality has been explored relatively rarely: Booth and Hux (2006) found that patients coming from neighbourhoods characterized by a low socioeconomic status (SES) are hospitalized more frequently; a significant effect of neighbourhood SES on hospitalization rates has been found also by Taylor et al. (2006), who also found a significant interaction between individual and neighbourhood SES. We prefer number of hospitalizations above number of visits to a physician in the last year and above perceived (self-assessed) health status, the main alternatives, because the former is a more objective measure of health status with respect to perceived health and also, to a lesser extent, with respect to the number of visits to a physician. In prospect, it may also be interesting to model both hospitalizations and visits in a bivariate setting. A further alternative measure of health status would be the fact of having been diagnosed as suffering from specific pathologies, such a choice, however, would be appropriate for a study devoted to specific pathologies rather than a study concerned with general health. Also, information on hospital admissions is more easily gathered (also using administrative sources), so the study can be more easily generalized. Characteristics investigated as determinants of human health can be related to demographic structure (age, gender and immigration structure), physical status (existing pathologies), lifestyle, socio-demographic situation (family structure and education attainment), socio-economic situa-

tion (level of poverty, income deprivation and residential stability), social resources availability (social cohesion and trust, information networks), environmental stressors (physical disorder and safety). Such investigations may be performed by computing neighbourhood level characteristics (for example by aggregating individual characteristics of inhabitants, like median income) or observing proper neighbourhood attributes (for example, pollution level) and associating them to each indivudal living in the area.

Rather than using neighbourhood level characteristics (quantities computed for each neighbourhood and then associated to each individual living in that neighbourhood), we prefer to consider the perception that the individual has of the neighbourhood. The fact of using individual perceptions about the neighbourhood may be seen as a limitation of the present study, they should in fact be regarded as measures with error of the neighbourhood quality. Using the perceptions, however, may also be a strength, since we avoid choosing a – forcefully arbitrary – definition of neighbourhood; moreover, it may be argued that the effect of the neighbourhood is not merely the result of its objective qualities, but also depends on subjective factors (this may be true in particular for social characteristics, but, even if we consider an eminently objective feature such as pollution, its effect on a subject depends on his personal degree of exposure and so is subjective). In particular, geographic or administrative boundaries – which are the most common choices – may not be the appropriate definition when dealing with characteristics related to social interactions (Diez Roux (2001)), in this case the 'perceived' neighbourhood (which is the definition implicitly used here) may be a more appropriate choice. The fact of using perception of the neighbourhood quality is one more reason to prefer hospitalization frequency over perceived health status as the response variable: using perceived health status any association to perceived neighbourhood quality may be spurious, since it may be driven by the overall attitude of the respondent.

The relationship between health and the perceptions about the neighbourhood has been already discussed in the literature. A variety of perceived aspects of the neighbourhood is considered by the various authors, broadly speaking one may distinguish perception of the physical (or environmental) characteristics of the neighbourhood and perceptions related to the social capital, the latter may concern problems due to (anti-)social behaviour on the negative side and to social cohesion on the positive side. Both the perceptions of physical and social characteristics were found to have a significant effect on health in Glasgow (UK) by Ellaway and Kearns (2001) and in Hamilton (CA) by Wilson et al. (2004). Subramanian et al. (2002) estimated a more complex relationship involving an interaction between individual trust and a contextual variable measuring community trust. As far as L.A.FANS data are concerned, Shin et al. (2006) found, estimating a logistic model, that social capital (as measured by the variables called closeknit, safe and neigh.satisf in Table 3) has a significant effect on self reported health in poor and very poor neighbourhoods.

2 Data

The Los Angeles Family and Neighbourhood Survey (L.A.FANS, Peterson et al. (2004)) is a panel study performed by the RAND corporation. In L.A.FANS a representative sample of households in Los Angeles County has been interviewed in two waves (2000-2001 and 2004-2005). The study is aimed at offering a better understanding of neighbourhood

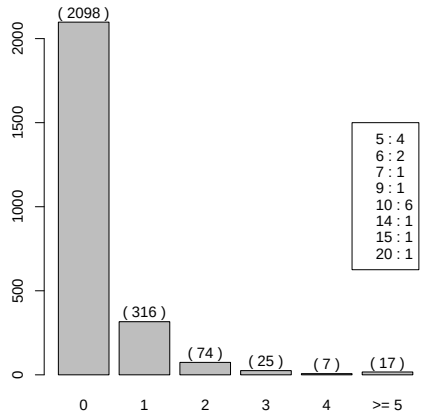

Figure 2: Barplot of the number of hospital admissions for RSA, in the box the frequencies of number of hospitalizations of RSA with 5 or more hospitalizations (in addition, 82 missing observations are present).

effects. For this reasons questionnaires ask for informations concerning neighbourhood characteristics and the random sample is stratified by neighbourhood – 65 neighbourhood (census tracts) in L.A. county were considered. Also, poor neighbourhoods and families with children are oversampled.

In the first wave, which is considered here, 3085 households were sampled (of which, 777 cases were households without children and 2308 with children) and for each of them interviews were made to a Randomly Selected Adult (RSA) and, if the household had children, a Randomly Selected Children (RSC), a Sibling (SIB) and the Primary Care Giver (PCG); in the end 2620 adults, 3161 children (2001 RSC and 1160 SIB) and 2044 caregivers completed the interview (see figure 2.3 and table 2.8 in Peterson et al. (2004)). In this work we do not consider family issues, so we consider only answers from the 2537 RSA interviewed who responded to the question concerning hospitalizations.

As already said the response variable we consider is the number of hospitalizations the respondent has incurred in the 24 months before interview (question AM25, see codebook by Peterson et al. (2004)). The frequency distribution of the response variable, number of hospitalizations in the last two years (Y_i in what follows), is depicted in Figure 2.

In Tables 1 and 2 we briefly describe the individual variables taken under consideration, distinguishing, for convenience, between numerical variables and factor variables. These are related to demographic structure (age, gender and immigration status), physical status (body mass index, previous diagnoses), socio-economic and socio-demographic situation (level of poverty and income deprivation, employment status, education level, marital status, having a regular source of care and residential stability).

In Table 3 we list variables describing the characteristics of the neighbourhood as perceived by the respondent (which are the only information on neighbourhood considered here). Such variables are related to social resources availability (social cohesion and trust, informal social control), and environmental stressors (safety) and general satisfaction with neighborhood. We outline in Table 3 a partition of the variables by topic, distinguishing informal social control, social cohesion, safety and general satisfaction with the neighbour-

hood. The responses on these topics are highly dependent, so it is advisable to build indicators which sum up subsets of them rather than to use all of them as covariates. A similar approach can be found in the literature: in Ellaway and Kearns (2001), Wilson et al. (2004), Subramanian et al. (2002), Shin et al. (2006) a number of questions are asked related to the perceived aspects of the neighbourhood and the answers are usually collapsed in a few indicators (among the cited studies only that by Wilson et al. (2004) is different on this respect in that respondents were asked open ended questions on likes and dislikes concerning the neighbourhood, also in this case, however, they were eventually collapsed in a few indicators). In this work, attempts have been made, unsuccessfully, at using data driven techniques such as cluster analysis to define indicators, eventually we defined indicators heuristically.

In particular, informal social control is measured by the first three variables in Table 3. We dichotomized the three variables assigning value 'likely' (1) to those units who responded 'very likely' or 'likely'; 'unlikely' (0) to those who responded 'unlikely' or 'very unlikely', the response 'unsure' is treated as a missing value. An indicator of informal social control is defined as the (not weighted) sum of the variables above. Variables concerning social cohesion have also been dichotomized after a preliminary analysis as 'agree' (corresponding to values 'strongly agree' and 'agree' in the original coding) or 'disagree' (corresponding to values 'disagree' or 'strongly disagree' in the original coding), treating as missing answer the value 'unsure'. An indicator for social cohesion and trust was defined as the (not weighted) sum of the 5 dichotomized variables concerning social cohesion. Eventually, safety is measured by the two variables safe and robbed, which have been used separately; it is to be noted that the latter is particularly interesting since it is the only one with an objective value (the other one is the perception about the safety in the neighbourhood), so it is worth using it separately.

Name	mean	min	max	s.d.	description
nhosp	0.33	0	20	1.13	Number of hospital admissions in the 24 months preceding interview
age	41.13	18	89	14.53	Age of respondent
bmi	26.25	17.48	40.39	4.17	Body mass index (BMI) of respondent
proincome	14157.08	0.00	190000.00	19875.43	Per capita income within family

Table 1: Relevant numerical variables in L.A.FANS. Body mass index is the ratio between self reported weight (in kg) and squared height (in meters); per capita income is given by the total income divided by the number of cohabitants (it is to be noted that this variable is censored since it is the sum of censored variables).

3 Methods

3.1 Background

Several studies of local area effects have been conducted focusing on the impact of neighbourhood (contextual) characteristics on health oucomes after adjusting for individual (compositional) status. The nature of the questions posed is such that one has to make use of statistical tools which allow for the multidimensionality of the phenomenon, the heterogeneity of its various aspects and the complexity of realities involved.

In this context, hierarchical regression analysis has become widely accepted as the appropriate tool for examining group level effects on individual-level health risks and outcomes: this kind of models allows for the nested structure of the data and thus allows for the variation between both individuals and groups. O'Campo (2003) discusses advancing theory and methods for multilevel models of residential neighbourhoods and health, see also Pickett and Pearl (2001) for a critical literature review on the statistical methods, the rationales and the quality of multilevel epidemiological studies on health published until 1998 (the oldest being published in 1983). Among more recent studies which make use of multilevel-hierarchical models we can find, for instance, those of Wen et al. (2003), Hou and Myles (2005), Larsen and Merlo (2005), Matheson et al. (2006), Stjärne et al. (2006), Dupéré and Perkins (2007), Fone et al. (2007). In all of them either standard logistic, ordinal logit or linear regression models are used according to the dichotomous or ordinal or continuous nature of the response variable. A hierarchical Poisson model is used when count variables are used, for instance in Petrelli et al. (2006) where incident cases of coronary heart disease are considered and in Aneshensel et al. (2007) where occurrence of depressive symptoms are analysed.

It is interesting to cite some of the existing publications that use or describe the L.A.FANS data analysed in this paper (see Section 2). The majority of them use the by now traditional methodological and statistical approach of multilevel regression analysis in order to conceptualize and measure how individual-level health risks and outcomes are affected by context (not only the neighbourhood of residence but, in some cases, also the so-called "activity space" that is the geography of the individual's daily activity over time). Examples of these are Dixon (2004), Inagami et al. (2006) and Inagami et al. (2007), Prentice (2006), Brown et al. (2007), Carpiano (2007), Frank et al. (2007).

More innovative alternatives to the by now traditional multilevel models in the literature concerned with urban and health data analysis include graphical models and spatial approaches. Rajulton and Niu (2005), thanks to the availability of longitudinal data, apply chain graph models, suitable for studying interrelationships among the structural, behavioral and health variables. Spatial analysis is considered in Glazier and Gozdyra (2004) and expecially Chaix et al. (2005, 2006, 2007). The latter, studying the prevalence of substance-related disorders, use, thanks to the availability of individual information geocoded at the exact place of residence, hierarchical geostatistical models, suitable for studying associations between contextual factors and health and for investigating not only the magnitude but also the scale of neighbourhood variations independent of administrative boundaries.

3.2 Models

Conceptually, we want to estimate an asymmetric model involving the number of hospitalizations (Y) as the response variable and a selection of determinants as the explanatory variables. In particular, we consider generalized additive models (GAM) since we want to be able to consider non linear effects of the explanatory variables on the response and since GAM allow analysing separately the effect of each covariate. Non linear contributions are estimated by spline functions whose degree of smoothness is decided by generalized cross validation.

Being Y a frequency, the traditional model for it would be based on the Poisson assumption

$$P(Y_i = y) = \frac{\mu_i^y}{y!} e^{-\mu_i} \tag{1}$$

and on the logarithmic link function between the parameter and the linear predictor, so that, if $x_{h,i}$, $h = 1, \ldots, H$, $i = 1, \ldots, n$ is the observed value of the h covariate on the i-th unit,

$$\log(\mu_i) = \alpha_0 + \sum_{h \in H_s} g_h(x_{h,i}) + \sum_{h \in H_l} \alpha_h x_{h,i}, \tag{2}$$

where $g_h(\cdot)$ are spline functions. The effects of covariates indexed in $H_s \subset \{1, \ldots, H\}$ are modeled non linearly, those indexed in $H_l \subset \{1, \ldots, H\}$ are modeled linearly ($H_s \cap H_l = \phi$ and $H_s \cup H_l \subset \{1, \ldots, H\}$).

The Poisson assumption may be too restrictive for real count data, a common extension is the Zero Inflated Poisson model (ZIP, Lam et al. (2006), Rigby and Stasinopulos (2005)), that is, one assumes that

$$Y_i = Z_i X_i, \tag{3}$$

where Z_i is 0 with probability π_i and 1 with probability $1 - \pi_i$ and $(X_i | Z_i = 1) \sim$ Poisson(λ_i), meaning that

$$P(Y_i = y) = \pi_i I(y = 0) + (1 - \pi_i) \frac{\lambda_i^y}{y!} e^{-\lambda_i}. \tag{4}$$

Covariates may affect either the parameter π_i and/or the parameter λ_i, through suitable link functions, the logistic and the logarithm functions respectively in this work, so we may get the two linear predictors

$$\text{logit}(\pi_i) = \beta_0^\pi + \textstyle\sum_{h \in H_s^\pi} g_h^\pi(x_{h,i}) + \sum_{h \in H_l^\pi} \beta_h^\pi x_{h,i}, \tag{5}$$

$$\log(\lambda_i) = \beta_0^\lambda + \textstyle\sum_{h \in H_s^\lambda} g_h^\lambda(x_{h,i}) + \sum_{h \in H_l^\lambda} \beta_h^\lambda x_{h,i}, \tag{6}$$

where all symbols are to be interpreted analogously to equation (2) ($H_s^\pi \cap H_l^\pi = \phi$ and $H_s^\lambda \cap H_l^\lambda = \phi$, while the other couples may have, pairwise, non empty intersections). Lam et al. (2006) recently proposed a method based on approximating the smooth functions by piecewise linear functions and on using the sieve maximum likelihood approach to obtain estimates to perform a semiparametric analysis within a ZIP assumption. We prefer the approach of Rigby and Stasinopulos (2005), which is based on the spline functions representation of smooth functions and the penalized likelihood approach to obtain estimates.

This latter approach is in fact, to our knowledge, more widely used and tested. In practice, estimation is made using the package `gamlss` (Stasinopulos and Rigby (2007)) in R (R Development Core Team (2005)).

The use of a ZIP model implies greater flexibility but, on the other side, interpretation of results is a bit harder. In fact, in a ZIP model the total effect of a variable, say u_i, on the expected number of hospitalizations, that is, the function $s(u) = E(Y|u, others)$, is a combination of its effects on the parameters λ and π, in fact

$$\begin{aligned} E(Y|\mathbf{x}) &= (1-\pi(\mathbf{x}))\lambda(\mathbf{x}) \\ &= \left(1 - \text{logit}^{-1}\left(\beta_0^\pi + \sum_{h\in H_s^\pi} g_h^\pi(x_h) + \sum_{h\in H_l^\pi} \beta_h^\pi x_h\right)\right) \times \\ &\qquad \times \exp\left(\beta_0^\lambda + \sum_{h\in H_s^\lambda} g_h^\lambda(x_h) + \sum_{h\in H_l^\lambda} \beta_h^\lambda x_h\right). \end{aligned} \tag{7}$$

The result of such a combination is not obvious and, moreover, the shape of the function $s(x_{h^*}) = E(Y|x_{h^*}, \mathbf{x}_{-h^*})$, also depends (due to the non linearity of the link functions) on the values of the other covariates. For this reason, in order to get a glance of the dependence of Y on x_h we compute $E(Y|x_{h^*}, \mathbf{x}_{-h^*})$ for a grid of values of the effect of x_{h^*}. In practice, we compute for all observed units i the quantities

$$v_i^\pi = \sum_{h\in H_s^\pi\setminus\{h^*\}} g_h^\pi(x_{h,i}) + \sum_{h\in H_l^\pi\setminus\{h^*\}} \beta_h^\pi x_{h,i}, \tag{8}$$

$$v_i^\lambda = \sum_{h\in H_s^\lambda\setminus\{h^*\}} g_h^\lambda(x_{h,i}) + \sum_{h\in H_l^\lambda\setminus\{h^*\}} \beta_h^\lambda x_{h,i}, \tag{9}$$

and consider, for some values $q \in [0,1]$, fixed empirical q-quantiles $v^\lambda_{([qn])}$ and $v^\pi_{([qn])}$ (where $[qn]$ is the integer part of qn). These values are then substituted in equation (7) to get

$$\begin{aligned} s_q^*(x_{h^*}) &= \left(1 - \text{logit}^{-1}\left(\beta_0^\pi + I_{H_s^\pi}(h^*) g_{h^*}^\pi(x_{h^*}) + I_{H_l^\pi}(h^*)\beta_{h^*}^\pi x_{h^*} + v^\pi_{([qn])}\right)\right) \times \\ &\quad \times \exp\left(\beta_0^\lambda + I_{H_s^\lambda}(h^*) g_{h^*}^\lambda(x_{h^*}) + I_{H_l^\lambda}(h^*)\beta_{h^*}^\lambda x_{h^*} + v^\lambda_{([qn])}\right), \end{aligned} \tag{10}$$

where $I_K(k)$ is an indicator function which is equal to one if $k \in K$ and to zero otherwise.

An alternative model one may consider is the Negative Binomial model, which is more flexible than the Poisson model (but not nested), the distribution of Y is then

$$P(Y_i = y) = \frac{\Gamma(y + 1/\sigma_i)}{\Gamma(y+1)\Gamma(1/\sigma_i)} \frac{(\mu_i\sigma_i)^y}{(\mu_i\sigma_i + 1)^{(y+(1/\sigma_i))}} \tag{11}$$

in which $\log(\sigma_i) = \gamma_0^\sigma$ and

$$\log(\mu_i) = \gamma_0^\mu + \sum_{h\in H_s^\mu} g_h^\mu(x_{h,i}) + \sum_{h\in H_l^\mu} \gamma_h^\mu x_{h,i} \tag{12}$$

is the linear predictor for μ.

Estimation is made, for all models, using the framework of Rigby and Stasinopulos (2005) implemented in package `gamlss` (Stasinopulos and Rigby (2007)) in R (R Development Core Team (2005)).

Our model selection strategy is first to choose the most relevant among physiological characteristics of the individual, then among the socioeconomic and sociodemographic ones and finally among those concerning the neighbourhood. Choice of candidate variables for inclusion is based on previous experiences accrued in the literature. Model comparison for variable selection is based on standard criteria: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and residual deviances comparison.

To assess model adequacy we check normality of the randomized residuals, that is, the quantities

$$r_i = (1 - u_i)\hat{F}_i(y_i - 1) + u_i\hat{F}_i(y_i), \tag{13}$$

where u_i are independent and identically distributed uniform random variables on $[0, 1]$, $\hat{F}$ is the Poisson, ZIP or NB distribution function with parameters equal to the estimated values (so for example in the ZIP case $\hat{F}_i(y) = F(y; \hat{\pi}_i, \hat{\lambda}_i)$ where F is the d.f. of the ZIP). In any case randomized residuals are, if the model is correct, normally distributed.

4 Results

In section 4.1 we report the results we got using ZIP model, in section 4.2 we briefly explore whether a model with a different distribution for the response variable – Poisson or negative binomial – fits well to the data.

4.1 ZIP model

In Table 4 we report the AIC, BIC and deviance values relevant for model comparisons. The procedure is incremental; in the i-th line we compare by AIC and BIC the current model, which involves all the variables selected up to and including step $i - 1$, and the model obtained applying the i-th modification: this is actually applied (and hence the current model for step $i + 1$ is the one resulting from the i-th modification) if it leads to a better AIC or BIC value than the current model and if the coefficient of the variable which is added is significantly different from zero at 5% level. It must be kept into account the fact that each variable has missing values for different units, so in order to do a fair comparison we must estimate both models, the current one and the modified one, on the same dataset (which is the largest one having no missing observations for the relevant variables). (It is to be noted that the inclusion of the quantities in Table 4 corresponds to the inclusion of a set of dummy variables when the quantity is a categorical variable; in this case the selection of significant coefficients may lead to the inclusion of only part of these dummy variables, that is, to an alternative definition of the factor levels.)

The model we start from (called $\mathcal{M}_0$ in what follows) includes the main physical characteristics: age, body mass index (BMI) and gender. In particular, the model $\mathcal{M}_0$ includes a non linear function of age and a dummy variable for gender in both linear predictors, while a non linear function of BMI enters only the linear predictor for π. In Figure 3 we report the non linear contributions of age and BMI. It is seen that the role of BMI is only slightly significant. The coefficient for the dummy variable for gender (valued one for males) is

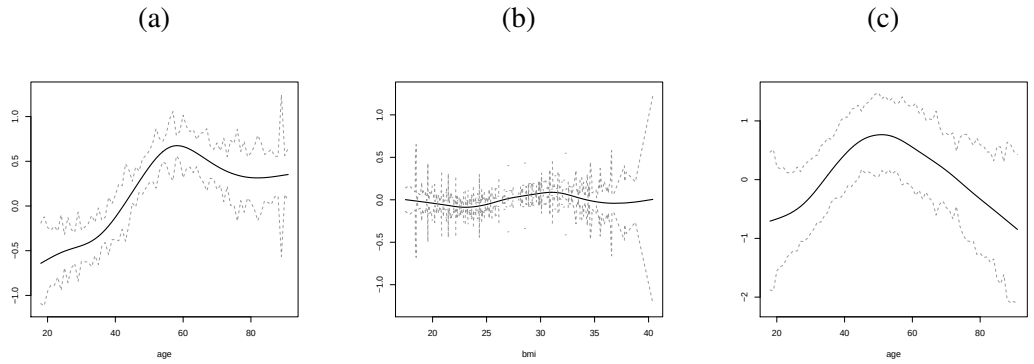

Figure 3: Non linear contributions to linear predictors for λ (panels (a) and (b)) and π (panel (c)) of the base model ($\mathcal{M}_0$).

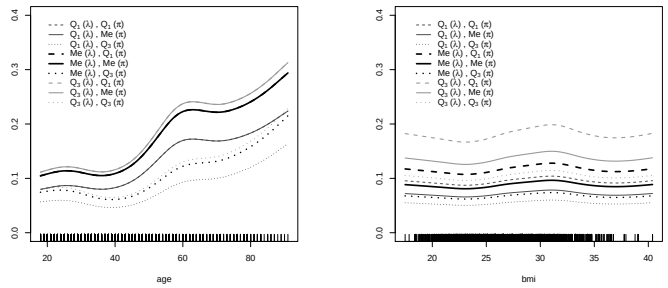

Figure 4: Estimated conditional expectation of Y according to the base model ($\mathcal{M}_0$).

-0.36749 (0.060307) in the linear predictor for λ and 0.8720 (0.146841) in the linear predictor for π, suggesting that on average males need hospitalization less frequently (this may have an obvious explanation since woman gets pregnant). In order to visualize the effect of age and BMI on Y we employ the technique explained in Section 3.2 obtaining the functions depicted in Figure 4. Age acts as anticipated, a higher number of hospitalizations is expected for older individuals. Also the effect of BMI is easily intepreted since both very low and very high values of BMI are associated with a higher number of hospitalizations (NIH (1998) classifies adults as underweight if BMI$<$ 18.5, normal if BMI range is 18.5-24.9, overweight if BMI range is 25.0-29.9 and obese if BMI $>$ 30). It must be kept in mind, however, that an abnormal value of the BMI may reasonably be either a risk factor or an intermediate factor associated with other pathologies (this is especially true for abnormally low values). Randomized residuals (Figure 7) suggest that the base model ($\mathcal{M}_0$) can be improved.

Below, we briefly comment on the reason of the inclusion for each variable (see Table 4), for a description of the variables see Section 2 and Tables 1, 2, 3.

Previous diagnoses (13 dummy variables) may be seen as intermediate causes between age and other personal characteristics and the number of hospitalizations in the causal chain.

We explore the effect of previous diagnoses by including in the model all of them ($\mathcal{D}_1$), only those which are less dependent on age ($\mathcal{D}_2$) and, finally, only those with estimated coefficient significantly different from zero at 5% ($\mathcal{D}_3$). Estimates of the model including all diagnoses confirm their intermediate role since the effect of age is reduced when the diagnoses are included. In the end, only chronic lung disease, excess weight and major depression are retained in the linear predictor for λ and only cancer/malignancy and emotional problems in the linear predictor for π.

Of the four dummy variables identifying the five ethnic groups of the variable **race** only 'White', 'Black', and 'Native/other' are significant in the linear predictor for λ; only 'White' in the l.p. for π.

After inclusion of the above variables (and of diagnoses in particular), estimates suggest simplifying the contribution of age by assuming a linear effect; such a choice leads to an improvement in terms of BIC (not of AIC).

As an indicator of wealth, we consider **per capita income** – household income divided by the number of components of the household (this is more representative of the money available to the individual rather than personal income). After exploring the inclusion of this variable as a linear or non linear function, we include a quadratic term for it only in the l.p. for λ. As is seen in Figure 5 and (more clearly) in Figure 6, the risk of hospitalization decreases as the income raises from low to average level, while it increases if the income is high. This may be due to non strictly necessary hospitalizations which are chosen by wealthy people (cosmetic surgery in Cook et al. (1995)): it is, however, to be noted that the uncertainty is high because few observations are available at the highest income levels; also, for privacy reasons the income variables which are summed up to compute per capita income are censored. Per capita income is kept into the model as a control variate even if it is only slightly significant when other variables are included in the model. Also, it is to be kept in mind that per capita income is related to the other socioeconomic characteristics which are also included: employment status, income from welfare and the fact of being a house owner.

Employment status is an indicator of wealth as well as per capita income (it is worth noting that employment would probably be collinear with personal income, this is not strictly the case with per capita income), but is also an indicator of lifestyle. It has a protective effect in the model.

The portion of income coming from **welfare** is kept separate from other incomes since it is an indicator of poorness rather than wealth (it would be improper to consider at the same wealth level two people having the same income, one from a job, the other from welfare); it is directly related to the number of hospitalizations with a significance of 10% , however, it is not significant at 5% level and so it is not included in the model.

The fact of being a **house** owner does not improve the model significantly, so it is not considered. Again, this information may be unimportant given the personal income.

Not surprisingly, the fact of having the highest level of **education** is associated with a lower expected number of hospitalizations as can be seen computing $E(Y)$ similar to what explained in equation (10). This characteristic is also potentially related to wealth.

The only significant variable related to **marital status** is the dummy for the fact of being neither married nor cohabitant; this has a positive coefficient in the l.p. for π, hence a protective effect.

The fact of having a **regular professional source of care** has no effect on the number of

hospitalizations (this is somewhat surprising since it was expected that such a characteristic would have been more frequently possessed by people with health problems).

Residential stability has a protective effect, those who moved in the neighbourhood since less than five years has a higher expected number of hospitalizations.

The perception of a more **dangerous neighbourhood** is associated with a higher mean of the number of hospitalizations, while the fact that a household has actually been **robbed** or has suffered a **vandalism** is non significant.

Eventually, a number of items is considered which measure the respondent perception of the neighbourhood.

Informal social control and **Social cohesion and trust** are introduced in the model either through two indicators or through two sets of dichotomized variables (see Section 2). Both features are not significant in model terms when indicators are used, on the contrary, if single items are considered, two of them, precisely 'getalong' and 'sharevalues' (Table 3) are signficant and both with a protective effect.

Neighbourhood satisfaction is measured by a categorical variable having 5 levels. This is either an indirect measure of the quality of the neighbourhood and a measure of the degree to which the respondent likes the neighbourhood he lives in. In both cases, we expect a low satisfaction to be associated with a higher number of hospitalizations. Model findings are difficult to interpret, a higher average number of hospitalizations is estimated for people answering 3 and 5 (which, however, constitute only 3% of the sample).

he model originated from the above selection procedure is then stripped of those variables which, despite leading to a lower AIC/BIC when included, have, in the last model, an estimated coefficient non significantly different from zero at 5% level. The resulting model is called $\mathcal{M}_1$ in what follows. The estimated coefficients of the linear components of $\mathcal{M}_1$ are reported in Table 5, Figure 5 depicts the non linear contributions to the linear predictors and Figure 6 depicts the effects of relevant variables on $E(Y)$, computed as explained in equation (10). In short, the final model evidenciates an overall monotonically increasing effect of age, which has an obvious interpretation. The effect of BMI is also easily interpreted on medical grounds since it is U shaped with a minimum corresponding to a normal weight-height ratio. (Low and high BMI may be a cause of illness or also a consequence of existing pathologies.) The risk of hospitalization decreases as the income raises from low to average level, while it increases if the income is high. This may be due to non strictly necessary hospitalizations which are chosen by wealthy people (cosmetic surgery in Cook et al. (1995)): it is, however, to be noted that the uncertainty is high because few observations are available at the highest income levels; also, for privacy reasons the income variables which are summed up to compute per capita income are censored.

Diagnostics for the model (right panel of Figure 7) show a satisfactory fit. In particular, the comparison of the normal probability plot of randomized residuals of the final model and that of residuals of the base model (left panel of Figure 7) clearly shows an improvement.

4.2 Comparison with Poisson and Negative Binomial models

In analysing count data as the number of hospitalization which is under consideration here the Poisson distribution is the most common modelling choice. When the Poisson model fails – as is easily seen to be the case here – alternatives include the Negative Binomial (NB) and the Zero Inflated Poisson (ZIP), whose appropriateness clearly depends on why

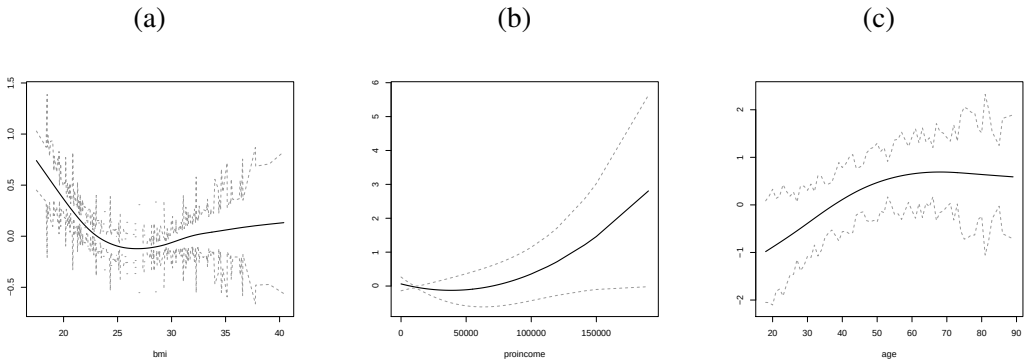

Figure 5: Non linear contributions to the linear predictors for λ (BMI in panel (a) and income in panel (b)) and π (age in panel (c)) of the final model ($\mathcal{M}_1$).

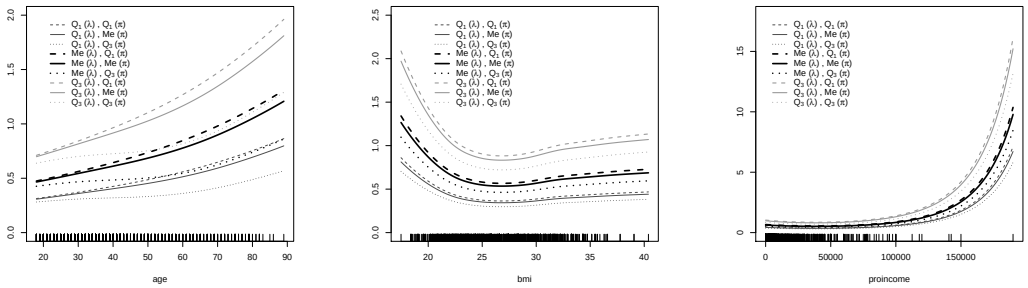

Figure 6: Estimated conditional expectation of Y according to the final model ($\mathcal{M}_1$).

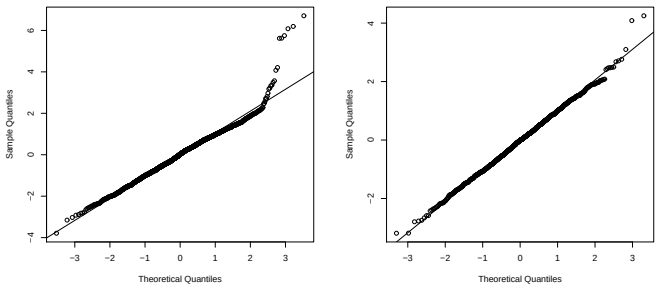

Figure 7: Normal probability plot of randomized residuals for the base model $\mathcal{M}_0$ (left) and the final model $\mathcal{M}_1$ (right).

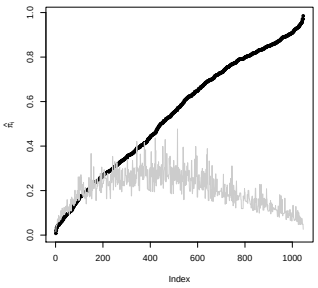

Figure 8: Sorted estimates $\hat{\pi}_i$ (dots) in model $\mathcal{M}_1$, gray line is $0 + 1.96 \times s.e.(\pi_i)$.

the Poisson model actually fails. In particular, the ZIP is an explicit attempt at modelling an excess of zero counts: it, in fact, postulates that the counts come from one of two regimes: either a degenerate component with unit mass in zero or a Poisson distribution. Being in presence of an excess of zero counts which can be fruitfully modelled separately is a relatively common situation in medical and public health statistics (Lam et al. (2006)). Such a model is also appealing on interpretation grounds since the probability of being in the zero-regime – which is explicitly estimated – may be interpreted as the probability of being in a lower risk group (subpopulation). (Clearly, this does not mean that one actually believes in the existence of two subpopulations.) The NB also allows for overdispersion with respect to a Poisson distribution but in a less specific way and does not share a similar interpretation.

It is worth noting that the ZIP distribution for Y_i differs from a Poisson distribution to the extent that π_i differs from 0 (since if $\pi_i = 0$ then $Y_i \sim \text{Poisson}(\lambda_i)$). We may then get a hint about the relevance of the ZIP model by checking how often $\hat{\pi}_i$ is significantly different from 0. We depict such a comparison in Figure 8 where we plot the estimates $\hat{\pi}_i$ ordered increasingly and, as a reference, the value of the standard error of each $\hat{\pi}_i$ multiplied by 1.96. Results show that $\hat{\pi}_i$ is, in prevalence, different from 0. This graphical display substantially resembles the kind of tests which are commonly suggested in the literature (van den Broek (1995); Rodrigues (2006)) where, usually, the hypotheses $H_0 : \pi = 0$ is tested against $H_1 : \pi \neq 0$ by means of a score test, whose main advantage is that one does not need to estimate the more complicated Poisson model. Jansakul and Hinde (2002) consider a model in which π_i is modeled as a function of the covariate ($\log(\pi/(1-\pi)) = X\gamma$) and propose a score test for the hypotheses $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$, rather than adapting the score test to the case of predictors with smooth function we prefer estimating both model and compare them in Table 6. A test for ZIP versus Poisson can be based, since they are nested models, on the difference between the deviances ($1484.8 - 1307 = 177, 8$) which, under $\mathcal{M}_2$, has a χ^2_{31-23} distribution, thus leading to a p-value which is almost equal to zero.

In considering the ZIP as a model alternative to the Poisson (to be considered a base model in this context) one should keep in mind, following El-Shaarawi (1985) but also Thas and Rayner (2005) that the mere preference for a ZIP over a Poisson model (that is, the fact that the lack of fit of the Poisson distribution is due to the excess of zeros) does not imply that the former is the appropriate choice, for this reason it is worth comparing the fit also against the NB.

In order to have a fair comparison of the ZIP, the Poisson ($\mathcal{M}_2$) and the NB ($\mathcal{M}_3$)

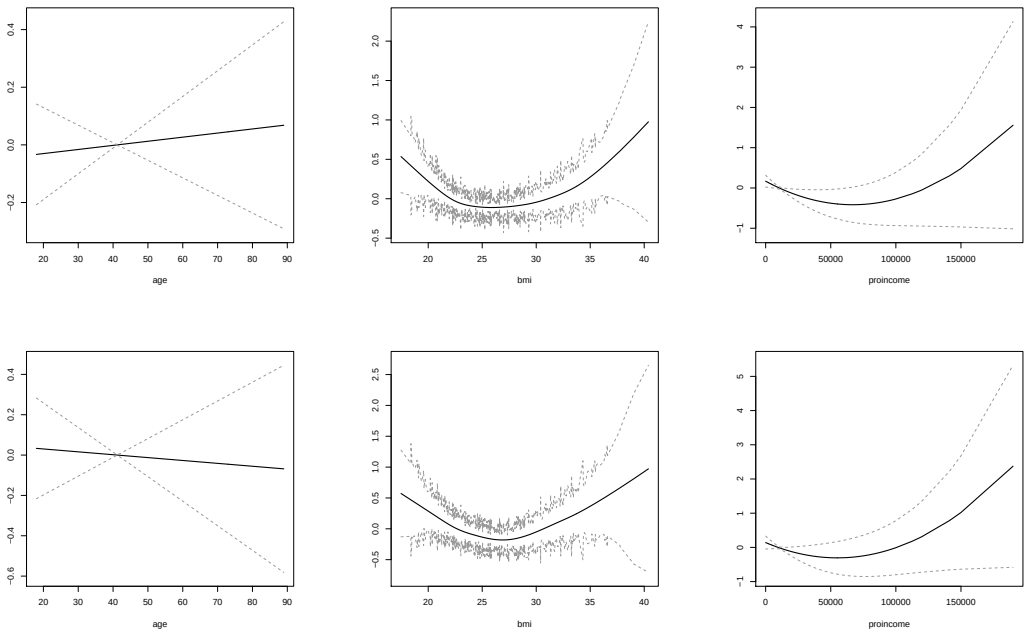

Figure 9: Contributions to the linear predictor for μ in the Poisson model $\mathcal{M}_2$ (top row) and in the Negative Binomial model $\mathcal{M}_3$ (bottom row).

model, we estimate both $\mathcal{M}_2$ and $\mathcal{M}_3$ using as explanatory variables all variables included in the final model (in any linear predictor, in fact the conditional distribution of Y in the ZIP model depends on both π and λ). Estimates of coefficients are in Table 5.

BMI and income, whose contributions to the linear predictors in Figures 9 and 6 are similar in shape, play an analogous role in the three models (the linear predictor in the Poisson and Negative Binomial models is directly related to $E(Y)$, so we can compare its shape with that of $E(Y)$ in the ZIP model). The contribution of age is non significantly different than a constant in the models $\mathcal{M}_2$ and $\mathcal{M}_3$ which may be the result of a lack of fit. The comparison of randomized residuals with a gaussian distribution (Figure 10) suggests lack of fit in the right tail for $\mathcal{M}_2$, the Negative Binomial and ZIP models leading to a significant improve in that area (Figure 7). Moreover, since the ZIP and Poisson model are nested, we can compare their AIC and BIC values, such a comparison (Table 6) is in favour of the ZIP model, as already confirmed by the significance of the test previously discussed.

Thus, diagnostics and tests are consistently in favour of the more flexible Negative Binomial and ZIP models over the Poisson model, as intuition suggested beforehand. The comparison between the ZIP and NB models (based on the residuals in Figure 7) leads to no clear cut conclusion.

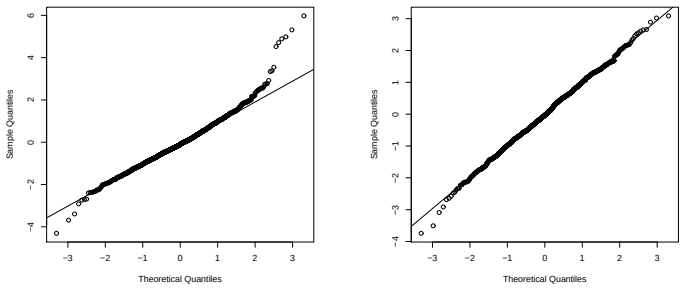

Figure 10: Normal probability plot of randomized residuals for the Poisson model $\mathcal{M}_2$ (left) and the Negative Binomial model $\mathcal{M}_3$ (right).

5 Final remarks

Using data from the L.A.FAN Survey we have investigated whether some neighbourhood characteristics, measured by the perception of the individual, have a significant effect on the number of hospitalizations. We have used generalized additive models which allow us to deal with non linear effects in a convenient way. We also overcome to some extent the difficulties in interpreting the results from a GAM with a ZIP distribution by simulating predicted values under varying assumptions in order to reveal the relationship of interest.

It turns out relatively clearly that a high social cohesion leads to a lower hospitalization rate as shown by the fact that both the variables 'getalong' and 'sharevalues' have significant coefficients with signs implying in both cases a negative effect on expected number of hospitalizations. Interestingly, the more general 'satisfaction' level concerning the neighbourhood does not show a clear effect.

Such results confirm the role of neighbourhood social environment on the health of individuals measured by the number of hospitalizations, particularly social cohesion which is, according to our results, more important than social control or public order. A strength of our conclusions is that they do not depend on a specific definition of neighbourhood since each respondent refers (implicitely) to his perception, on the other hand, this is also a limitation in that it does not allow to define 'good' and 'bad' neighbourhood. It is worth to stress that the design of the study does not allow to infer a causal relationship, on the contrary it is certainly possible that a 'bad' perception of the neighbourhood is caused by a poor health status (and, consequently, relatively frequent hospitalizations).

The comparison between Poisson, Negative Binomial and ZIP models shows that the performance of the traditional Poisson distributional assumptions for count data can be greatly improved, our results are not conclusive on whether a Negative Binomial or a ZIP model is to be preferred, the latter, however is more flexible and so is preferred here. The choice of the model is relevant for our purposes since different models lead to different conclusions on the significance of the covariates (as seen in Table 5). For instance, age, education and some of the diagnoses are not significant neither for the Poisson nor fot the Negative Binomial models while they are significant (as expected) in the ZIP model.

Acknowledgements

The authors are supported by University of Padua (grant n. 062874). This research is based on data from the Los Angeles Family and Neighborhood Survey which is funded by a grant R01 HD35944 from the National Institute of Child Health and Human Development to RAND in Santa Monica, California. For further information on L.A.FANS, go to `www.lasurvey.rand.org`. We would like to thank professor Gianfranco Lovison (University of Palermo, Italy) for his comments.

Name	num. levels	description (levels, 1-level for dichotomic)
gender	2	gender (male)
race	5	race (latino, white, black, asian/pacific, native/other)
empl	4	employment status (unemployed (not recent), never employed, unemployed (recent), employed)
welfareinc	2	income from welfare (yes)
house	3	status house (owned, rented, other)
edu	3	education (less, high, college)
marstatus	4	marital status (married, living with partner, neither, both)
rsoc	2	having a regular source of care different than relatives and friends (yes)
res.stab	2	residential stability (moved since less than 5 years)
		Previous diagnoses[a]
cld	2	diag. of chronic lung disease
excweight	2	diag. of excess weight
depress	2	diag. of major depression
cancer	2	diag. of cancer/malignancy
emotional	2	diag. of emotional problems
hbp	2	diag. of high blood pressure
diabetes	2	diag. of diabetes
ha	2	diag. of heart attack
chd	2	diag. of coronary heart disease
artrrheum	2	diag. of arthritis or rheumatisms
asthma	2	diag. of asthma
mentalloss	2	diag. of loss of mental ability
learndis	2	diag. of learning disorder

[a]Diagnoses variables are non null if the interviewed answered positively to the question 'Have you ever been diagnosed ...'.

Table 2: Relevant non numerical variables in L.A.FANS database.

Name	num. levels	description (levels)
		Variables concerning informal social control[a]
hangout	2	neighbours do something if kid hangs out
graffiti	2	would do something if kid does graffiti
disresp	2	would scold kid if showing disrespect
inf.soc.ctrl		index of informal social control[b]
		Variables concerning social cohesion[c]
closeknit	2	this is a close knit neighbourhood
willhelp	2	people are willing to help neighbours
getalong	2	neighbours generally get along[d]
sharevalues	2	people in neighbourhood share same values[e]
trusted	2	people in neighborhood can be trusted
cohes.trust		index of social cohesion and trust[f]
		Variables concerning safety in neighborhood
safe	4	how safe is to walk around alone (1=completely safe to 4=extremely dangerous)
robbed	2	household has been robbed or suffered vandalism in the neighbourhood (0=no to 1=yes)
		Other variables describing neighborhood
neigh.satisf	5	degree of satisfaction with the neighborhood (1=very satisfied to 5=very dissatisfied)

[a] All but the last are coded as 1='very likely'; 2='likely'; 3='unsure'; 4='unlikely'; 5='very unlikely' in the original coding, they are dichotomized as 'unlikely'=0='unlikely' or 'very unlikely'; 'likely'=1='very likely' or 'likely'; response 'unsure' was ignored as a missing value.

[b] The (non weighted) sum of the variables concerning informal social control as measured in the orginal scale.

[c] All but the last are coded as 1='strongly agree'; 2='agree'; 3='unsure'; 4='disagree'; 5='strongly disagree' in the original coding, they are dichotomized as 'Agree'=1='strongly agree' or 'agree'; 'Disagree'=0='disagree' or 'strongly disagree'; response 'unsure' was ignored as a missing value.

[d] Variable has been given the opposite sense here than in L.A.FANS questionnaire (original question was 'neighbors generally don't get along').

[e] Variable has been given the opposite sense here than in L.A.FANS questionnaire (original question was 'people in neighbourhood don't share same values').

[f] The (non weighted) sum of the variables concerning social cohesion and trust as measured in the orginal scale but for the transformations in notes *d* and *e* above.

Table 3: Relevant non numerical variables related to neighborhood in L.A.FANS database.

Cathegories in the CF				AIC		BIC		deviance	
		n	d.f.	w/out	with	w/out	with	w/out	with
Physical characteristics	**+g(age)+gender+g(BMI)**	2343	16	–	3061.12	–	3153.3	–	3029.1
	$+D_1$	2336	35	3058.63	2868.83	3150.7	3070.3	3026.6	2798.8
	$+D_2$	2336	28	2868.83	2992.48	3070.3	3153.6	2798.8	2936.5
	$+\mathbf{D_3}$	2336	21	2992.48	2989.20	3153.6	3110.1	2936.5	2947.2
	+race	2336	25	2989.20	2920.66	3110.1	3064.6	2947.2	2870.7
	-g(age)+age	2336	21	2920.66	2925.20	3064.6	3046.1	2870.7	2883.2
Socioeconomic characteristics	**+g(proincome)**	2295	23	2875.82	2870.47	2996.3	3002.5	2833.8	2824.5
	+empl	2295	25	2870.47	2819.70	3002.5	2963.2	2824.5	2769.7
	+welfareinc	2283	26	2797.18	2797.15	2940.5	2946.2	2747.2	2745.2
	+house	2291	29	2818.09	2820.50	2961.5	2986.9	2768.1	2762.5
Sociodemographic characteristics	**+edu**	2265	27	2791.33	2790.53	2934.5	2945.1	2741.3	2736.5
	+marstatus	2265	28	2790.53	2777.46	2945.1	2937.8	2736.5	2721.5
	+rsoc	2261	29	2776.30	2774.16	2936.6	2940.1	2720.3	2716.2
Neighbourhood characteristics	**+res.stab**	1712	29	2120.43	2119.75	2272.9	2277.7	2064.4	2061.7
	+safety	1450	30	1908.62	1899.82	2061.7	2058.2	1850.6	1839.8
	+inf.soc.ctrl	1302	31	1690.83	1689.96	1846.0	1850.3	1630.8	1628.0
	+cohes.trust	1118	31	1503.48	1504.58	1654.1	1660.2	1443.5	1442.6
	+i.s.c & c.t	1046	32	1404.09	1397.59	1552.7	1556.1	1344.1	1333.6
	+neigh.satisf	1046	40	1397.59	1380.00	1556.1	1578.1	1333.6	1300.0
	$-I$	1046	31	1380.00	1368.99	1578.1	1522.5	1300.0	1307.0

$\mathcal{D}_1$: set of all diagnoses variables (see Table **??**);
$\mathcal{D}_2$: set of those diagnoses which are less correlated with age (excess weight, depression, cancer, emotional disorder, chronic lung disease, asthma, loss of mental ability, learning disorder);
$\mathcal{D}_3$: set of those diagnoses which are significant in the model (that is, excess weight, depression, cancer, emotional disorder and chronic lung disease);
-g(age)+age stands for the replacement of the non linear contribution of age with a linear one;
i.s.c & c.t stands for the set of the dichotomized variables related to informal social control and to social cohesion amd trust (listed in Table 3) having an estimated coefficient significantly different from 0 at 5%;
I is the set of variables whose coefficients are not significantly non null.

Table 4: Pairwise comparisons of models for successive additions of variables from the base model $\mathcal{M}_0$ (including age, gender and BMI) to the final model $\mathcal{M}_1$ (including variables in bold face in the table).

Cathegories in the CF		$\mathcal{M}_1$ λ	$\mathcal{M}_1$ π	$\mathcal{M}_2$ μ	$\mathcal{M}_3$ μ
	(Intercept)	-0.405	-2.467$'''$	-0.510	-0.221
		(0.4958)	(0.7082)	(0.4187)	(0.5948)
Physical characteristics	age	0.015$'''$	-	0.001	-0.001
		(0.0044)	-	(0.0038)	(0.0054)
	gender=male	-0.209	-	-0.083	-0.322
		(0.1868)	-	(0.1333)	(0.1807)
	cld	0.434$''$	-	0.644$'''$	0.704$'$
		(0.1506)	-	(0.1811)	(0.3019)
	excweight	0.516$'''$	-	0.272	0.207$'$
		(0.1490)	-	(0.1461)	(0.2148)
	depress	0.480$''$	-	0.915$'''$	1.071$'''$
		(0.1541)	-	(0.1939)	(0.2954)
	cancer	-	-1.645$'$	0.313	0.637
		-	(0.7035)	(0.2100)	(0.3274)
	emotional	-	-0.909$'$	-0.218	-0.130
		-	(0.4145)	(0.2098)	(0.3105)
	race=native/other	2.767$'''$	-	2.364$'''$	2.184$'''$
		(0.1629)	-	(0.2276)	(0.5125)
Socioeconomic characteristics	empl=employed	-	1.766$'''$	-1.151$'''$	-1.122$'''$
		-	(0.3083)	(0.1281)	(0.1656)
Sociodemographic characteristics	edu=college	-1.161$'''$	-2.249$'''$	-0.246	0.009
		(0.2188)	(0.4855)	(0.1645)	(0.2068)
	marstatus=neither	-	0.691$''$	-0.302$''$	-0.447$''$
		-	(0.2580)	(0.1163)	(0.1617)
Neighbourhood characteristics	getalong	-0.405$'$	-	-0.378$''$	-0.381$'$
		(0.1623)	-	(0.1295)	(0.1826)
	sharevalues	-	0.727$''$	-0.152	-0.352$'$
		-	(0.2698)	(0.1179)	(0.1607)
	neigh.satisf=2	0.169	0.391	-0.083	0.024
		(0.1941)	(0.3648)	(0.1491)	(0.2014)
	neigh.satisf=3	1.804$'''$	2.080$''$	0.305	0.309
		(0.2724)	(0.7965)	(0.3037)	(0.4374)
	neigh.satisf=4	0.057	0.482	-0.304	-0.256
		(0.2608)	(0.4879)	(0.2071)	(0.2815)
	neigh.satisf=5	1.284$'''$	1.098$'$	0.582$''$	0.5702
		(0.2522)	(0.5405)	(0.2186)	(0.3491)

Table 5: Estimated coefficients (and their standard erros in parenthesis) in the linear predictors according to models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ ($'$ denotes significance at 0.05, $''$ at 0.01, $'''$ at 0.001).

	d.f.	AIC	BIC	deviance
$\mathcal{M}_1$	31	1368.99	1522.5	1307.0
$\mathcal{M}_2$	23	1530.76	1644.7	1484.8
$\mathcal{M}_3$	24	1370.99	1489.8	1323.0

Table 6: Comparison of ZIP, Poisson and Negative Binomial models: $\mathcal{M}_1$ is the final ZIP model; $\mathcal{M}_2$ and $\mathcal{M}_3$ are, respectively, the Poisson and NEgative Binomial models involving all variables included in $\mathcal{M}_1$.

References

Aneshensel, C., Wight, R., Miller-Martinez, D., Botticello, A., Karlamangla, A., Seeman, T., 2007. Urban neighborhoods and depressive symptoms among older adults. J Gerontol B Psychol Sci Soc Sci 62, S52–S59.

Bellini, P., Braga, M., Rebba, V., Rodella, S., Vendrami, E., 2002. Definizione di un set di indicatori per il monitoraggio e la valutazione dell'attività sanitaria. Tech. Rep. 02.03, CGIS, Commissione per la Garanzia dell'Informazione Statistica, Roma.
URL www.palazzochigi.it/Presidenza/Statistica/attivita/rapporti/2002/02.03.pdf

Bellini, P., Rodella, S., May-June 2004. Indicatori di qualità dei servizi socio-sanitari. La definizione di un modello concettuale. Salute e Territorio, 169–179.

Berkman, L., Kawachi, I. (Eds.), 2000. Social epidemiology. Oxford University Press.

Booth, G., Hux, J., 2006. Relationship between avoidable hospitalizations for diabetes mellitus and income level. Arch Intern Med 163, 101–106.

Brackstone, G., 1987. Issues in the use of administrative records for statistical purposes. Survey Methodology 13, 29–43.

Brown, A., Ang, A., Pebley, A., 2007. The relationship between neighborhood characteristics and self-rated health for adults with chronic conditions. American Journal of Public Health 97, 926–932.

Canadian Institute for Health Information, 1999. National consensus conference on population health indicators. Final report.

Carpiano, R., 2007. Neighborhood social capital and adult health: An empirical test of a Bourdieu-based model. Health & Place 13, 639–655.

Chaix, B., Leyland, A., Sabel, C., Chauvin, P., Rstam, L., Kristersson, H., Merlo, J., 2006. Spatial clustering of mental disorders and associated characteristics of the neighbourhood context in Malmö, Sweden, in 2001. J Epidemiol Community Health 60, 427–435.

Chaix, B., Merlo, J., Subramanian, S., Lynch, J., Chauvin, P., 2005. Comparison of a spatial perspective with the multilevel analytical approach in neighborhood studies: The case of mental and behavioral disorders due to psychoactive substance use in Malmö, Sweden, 2001. Am J Epidemiol 162, 171–182.

Chaix, B., Rosvall, M., Merlo, J., 2007. Assessment of the magnitude of geographical variations and socioeconomic contextual effects on ischaemic heart disease mortality: a multilevel survival analysis of a large Swedish cohort. J Epidemiol Community Health 61, 349–355.

Cook, R. R., Delongchamp, R. R., Woodbury, M., Perkins, L. L., Harrison, M. C., 1995. The prevalence of women with breast implants in the United States - 1989. Journal of Clinical Epidemiology 48, 519–525.

Diez Roux, A., 2001. Investigating neighborhood and area effects on health. American Journal of Public Health 91, 1783–1789.

Dixon, E., 2004. Neighborhoods and Adult Health Status: A Multi-Level Analysis of Social Determinants of Health Disparities in Los Angeles County. Doctoral Dissertation, UCLA, Los Angeles, CA.

Dupéré, V., Perkins, D., 2007. Community types and mental health: a multilevel study of local environmental stress and coping. Am J Community Psychol 39, 107–119.

ECHI, 2004. ECHI and conceptual schemes/models of health.
URL http://ec.europa.eu/health/ph_information/implement/wp/indicators/docs/ev_20040219_rd02_en.pdf

El-Shaarawi, A., 1985. Some goodness-of-fit methods for the Poisson plus added zeros distribution. Applied and environmental microbiology, 1304–1306.

Ellaway, A. ad Macintye, S., Kearns, A., 2001. Perceptions of place and health in socially contrasting neighbourhoods. Urban Studies 38, 2299–2316.

Fone, D., Dunstan, F., Lloyd, K., Williams, G., Watkins, J., Palmer, S., 2007. Does social cohesion modify the association between area income deprivation and mental health? a multilevel analysis. International Journal of Epidemiology 36, 338–345.

Frank, R., Cerdá, M., Rendón, M., 2007. Barrios and burbs: Residential context and health-risk behavior among Angeleno adolescents. J Health Soc Behav 48, 282–300.

Glazier, R., Gozdyra, P., 2004. Using Spatial Analysis and Maps to Understand Patterns of Health Services Utilization. From "Enhancing Information and Methods for Health System Planning and Research", Institute for Clinical Evaluative Sciences (ICES), January 19-20, Toronto, Canada.
URL http://www.torontohealthprofiles.ca/documents/resources/Presentation_ICES_workshop_12012004.pdf

Hancock, T., Duhl, L., 1998. WHO Healthy Cities project: a guide to assessing Healthy Cities. FADL Publishers, Copenhagen.

Hancock, T., Labonte, R., Edwards, R., 1999. Indicators that count! Measuring population health at the community level. Canadian Journal Public HealthSuppl 1.

Hou, F., Myles, J., 2005. Neighbourhood inequality, neighbourhood affluence and population health. Social Science & Medicine 60, 1557–1569.

Inagami, S., Cohen, D., Finch, B., 2007. Non-residential neighborhood exposures suppress neighborhood effects on self-rated health. Social Science & Medicine 65, 1779–1791.

Inagami, S., Cohen, D., Finch, B., Asch, S., 2006. You are where you shop: Grocery store locations, weight, and neighborhoods. American Journal of Preventive Medicine 31, 10–17.

Jansakul, N., Hinde, J., 2002. Score tests for zero-inflated Poisson models. Computational Statistics & Data Analysis 40, 75–96.

Lam, K., Xue, H., Cheung, Y., 2006. Semiparametric analysis of zero-inflated count data. Biometrics 62, 996–1003.

Larsen, K., Merlo, J., 2005. Appropriate assessment of neighborhood effects on individual health: Integrating random and fixed effects in multilevel logistic regression. Am J Epidemiol 161, 81–88.

Li, Y., Schneider, J., Bennett, D., 2007. Estimation of the mediation effect with a binary mediator. Statistics in medicine 26, 3398–3414.

Matheson, F., Moineddin, R., Dunn, J., Creatore, M., Gozdyra, P., R.H., G., 2006. Urban neighborhoods, chronic stress, gender and depression. Social Science & Medicine 63, 2604–2616.

McDowell, I., Spasoff, R., Kristjansson, B., 2004. On the classification of population health measurements. Am J Public Health, 388–393.

Murray, C., Ezzati, M., Lopez, A., Rodgers, A., Hoorn, S., 2003. Comparative quantification of health risks: conceptual framework and methodological issues. Popul Health Metr 1, 1–20.

National Centre for Health Statistics, 2003. Summary Measures of Population Health. Report on Findings on Methodological and Data Issues. FADL Publishers, Copenhagen.

NIH, 1998. Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults. The Evidence Report. National Institutes of Health - National Heart, Lung, and Blood Institute in cooperation with The National Institute of Diabetes and Digestive and Kidney Diseases.
URL `http://www.nhlbi.nih.gov/guidelines/obesity/ob_gdlns.pdf`

O'Campo, P., 2003. Invited commentary: Advancing theory and methods for multilevel models of residential neighborhoods and health. Am J Epidemiol 157, 9–13.

Peterson, C. E., Sastry, N., Pebley, A. R., Ghosh-Dastidar, B., Williamson, S., Lara-Cinisomo, S., 2004. The Los Angeles Family and Neighborhood Survey Codebook. RAND Corporation.

Petrelli, A., Gnavi, R., Martinacci, C., Costa, G., 2006. Socioeconomic inequalities in coronary heart disease in Italy: A multilevel population-based study. Social Science & Medicine 63, 446–456.

Pickett, K., Pearl, M., 2001. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. J Epidemiol Community Health 55, 111–122.

Prentice, J., 2006. Neighborhood effects on primary care access in Los Angeles. Social Science & Medicine 62, 1291–1303.

R Development Core Team, 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Rajulton, F., Niu, J., 2005. Health over the life course: A chain graph model of interrelationships among socio-demographic, societal and lifestyle factors. From "XXV International Population Conference", International Union for the Scientific Study of Population (IUSSP), July 18-23, Tours, France.
URL http://iussp2005.princeton.edu/download.aspx?submissionId=52392

Rigby, R. A., Stasinopulos, D., 2005. Generalized additive models for location, scale and shape. Applied Statistics 54, 507–554.

Rodella, S., Bellini, P., Braga, M., V., R., 2003. Measuring and comparing performance of health services: a conceptual model to support selection and validation of indicators. Draft version available from stefania.rodella@arsanita.toscana.it.

Rodrigues, J., 2006. Full Bayesian significance test for zero-inflated distributions. Communications in Statistics - Theory and methods 35, 299–307.

Shin, M., Clark, W., Maas, R., 2006. Social capital, neighborhood perceptions and self-rated health: Evidence from the Los Angeles Family and Neighborhood Survey (LAFANS). Working Paper CCPR-039-06, California Center for Population Research, UCLA.
URL www.ccpr.ucla.edu/ccprwpseries/ccpr_039_06.pdf

Starfield, B., 2002. Equity in health. Epidemiol Community Health 56, 483–484.

Stasinopulos, D., Rigby, R. A., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software 23.
URL http://www.jstatsoft.org/v23/i07

Stjärne, M., Fritzell, J., Ponce De Leon, A., Hallqvist, J., the SHEEP Study Group, 2006. Neighborhood socioeconomic context, individual income and myocardial infarction. Epidemiology 17, 14–23.

Subramanian, S., Kim, D., I., K., 2002. Social trust and self-rated health in US communities: a multilevel analysis. Journal of Urban Health 79, S21–S34.

Taylor, C., David, A., Winkleby, M., 2006. Neighborhood and individual socioeconomic determinants of hospitalization. American Journal of Preventive Medicine 31, 127–134.

Thas, O., Rayner, J., 2005. Smooth tests for the zero-inflated poisson distribution. Biometrics 61, 808–815.

van den Broek, J., 1995. A score test for zero inflation in a Poisson distribution. Biometrics 51, 738–743.

Webster, P. e. a., 1996. Healthy Cities Indicators: Analysis of Data from Cities across Europe. WHO, Copenhagen.

Wen, M., Browning, C., Cagney, K., 2003. Poverty, affluence, and income inequality: neighborhood economic structure and its implications for health. Social Science & Medicine 57, 843–860.

WHO Commission on Social Determinants of Health, 2005. Towards a conceptual framework for analysis and action on the social determinants of health. Discussion paper for the commission on social determinants of health.
URL www.who.int/social_determinants/resources/framework.pdf

WHO Commission on Social Determinants of Health, 2007. A conceptual framework for action on the social determinants of health. Discussion paper for the commission on social determinants of health.
URL www.who.int/social_determinants/resources/csdh_framework_action_05_07.pdf

Wilson, K., Elliott, S., Law, M., Eyles, J., Jerrett, M., Keller-Olaman, S., 2004. Linking perceptions of neighbourhood to health in Hamilton, Canada. J. Epidemiology and Community Health 58, 192–198.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it

Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy